

PI 1111330

RECEIVED

09 JAN 2004

WIPO

PCT

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office

January 06, 2004

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM
THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK
OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT
APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A
FILING DATE.

APPLICATION NUMBER: 60/477,314

FILING DATE: June 09, 2003

RELATED PCT APPLICATION NUMBER: PCT/US03/34028

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS



P. SWAIN
Certifying Officer

**PRIORITY
DOCUMENT**

SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

BEST AVAILABLE COPY

06-11-03

PTO/SB/16 (10-01)


Approved for use through 10/31/2002 OMB 0651-0037
Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PROVISIONAL APPLICATION FOR PATENT COVER SHEET
This is a request for filing a PROVISIONAL APPLICATION under 37 CFR 1.53(c).

Express Mail Label No. EV195881645US

Docket Number: 910000-2139

INVENTOR(S)					
Given Name (first and middle [if any])		Family Name or Surname		Residence (City and either State or Foreign Country)	
Keith J.		Joung		1 Magnolia Way Winchester, MA	
<input type="checkbox"/> Additional inventors are being named on the <u> </u> separately numbered sheets attached hereto					
TITLE OF THE INVENTION (500 characters max.)					
METHOD FOR PRODUCING HIGHLY SPECIFIC DESIGNER ZINC FINGER PROTEINS CAPABLE OF RECOGNIZING EXTENDED DNA SEQUENCES					
Direct all correspondence to:			CORRESPONDENCE ADDRESS		
<input type="checkbox"/> Customer Number <u>20999</u> OR Type Customer Number here			 20999 PATENT TRADEMARK OFFICE		
<input checked="" type="checkbox"/> Firm or Individual Name		Thomas J. Kowalski, Esq. FROMMER LAWRENCE & HAUG LLP			
Address		745 Fifth Avenue			
City	New York	State	New York	Zip	10151
County	New York	Telephone	(212) 588-0800	Fax	(212) 588-0500
ENCLOSED APPLICATION PARTS (Check all that apply)					
<input checked="" type="checkbox"/> Specification		Number of Pages		81 pages (Specification 80 pgs & 1 pg. Abstract)	
<input checked="" type="checkbox"/> Drawing(s)		Number of Sheets		35 sheets (Figures 1-35)	
<input type="checkbox"/> Application Data Sheet. See 37 CFR 1.76				<input type="checkbox"/> CD(s), Number	
				<input type="checkbox"/> Other (specify)	
METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT					
<input checked="" type="checkbox"/> Applicant claims small entity status. See 37 CFR 1.27. <input checked="" type="checkbox"/> A check or money order is enclosed to cover the filing fees <input type="checkbox"/> The Commissioner is hereby authorized to charge filing fees and credit Deposit Account Number: 50-0320 <input type="checkbox"/> Payment by credit card. Form PTO-2038 is attached.					FILING FEE AMOUNT (\$) \$80.00
The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.					
<input checked="" type="checkbox"/> No <input type="checkbox"/> Yes, the name of the U.S. government agency and the Government contract number are:					

Respectfully submitted,

SIGNATURE

Amy Leahy

DATE

June 9, 2003

TYPED OR PRINTED NAME

Thomas J. Kowalski, Esq.
Amy Leahy, Ph.D.

REGISTRATION NO.
(if appropriate)
Docket Number:

32,147
47,739

TELEPHONE

(212) 588-0800

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

This collection of information is required by 37 CFR 1.51. The information is used by the public to file (and by the PTO to process) a provisional application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 8 hours to complete, including gathering, preparing, and submitting the complete provisional application to the PTO. Time will vary depending upon the individual case. Any comments on the amount of time you are required to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, Patent and Trademark Office, U.S. Department of Commerce, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO Mail Stop Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

00131307

TITLE OF THE INVENTION

**METHODS FOR PRODUCING SPECIFIC DESIGNER ZINC FINGER PROTEINS
CAPABLE OF RECOGNIZING EXTENDED DNA TARGET SEQUENCES**

RELATED APPLICATIONS/PATENTS & INCORPORATION BY REFERENCE

Each of the applications and patents cited in this text, as well as each document or reference cited in each of the applications and patents (including during the prosecution of each issued patent; "application cited documents"), and each of the PCT and foreign applications or patents corresponding to and/or claiming priority from any of these applications and patents, and each of the documents cited or referenced in each of the application cited documents, are hereby expressly incorporated herein by reference, and may be employed in the practice of the invention. More generally, documents or references are cited in this text, either in a Reference List before the claims, or in the text itself; and, each of these documents or references ("herein cited references"), as well as each document or reference cited in each of the herein cited references (including any manufacturer's specifications, instructions, etc.), is hereby expressly incorporated herein by reference.

**STATEMENT OF RIGHTS TO INVENTION MADE UNDER
FEDERALLY SPONSORED RESEARCH**

This work was supported, in part, by the government by a grant from the National Institute of Health and the National Institute of Diabetes and Digestive and Kidney Diseases (K08 DK02883). The government may have certain rights to this invention.

FIELD OF THE INVENTION

The present invention relates to multi-finger Zinc finger polypeptides that bind to extended DNA target sequences, and methods of selection thereof.

BACKGROUND

At any given time, only a fraction of the genes in the genome of an organism are expressed and/or producing functional protein products. The profile of proteins expressed in an organism varies greatly between cell types and changes over time, depending on factors such as

stage of development, stage of the cell cycle and response to environmental factors.

Furthermore, gene expression is often mis-regulated in disease.

Gene expression is controlled, in part, by proteins known as transcription factors. The presence of a particular combination of such transcription factors determines whether a gene is switched on or off at any given time and place. Transcription factors are modular proteins. They contain at least one DNA-binding domain (DBD) and one or more effector or regulatory domains. DBDs act as targeting devices to localize transcription factors to specific sequences or "target sites" on the chromosomal DNA. Effector domains function to direct the localization of specific activities to a gene or locus of interest, ultimately enabling transcription of that gene to be up- or down regulated.

The ability to artificially manipulate gene expression has enormous potential for biological research and for the development of new agents for gene therapy. Realizing this potential requires the ability to engineer DNA binding domains that recognize "target site" sequences with high affinity and specificity. Many DNA-binding proteins contain independently folded domains for the recognition of DNA, and these domains in turn belong to a large number of structural families, such as the leucine zipper, the "helix-turn-helix" and zinc finger (Zf) families. Most sequence-specific DNA-binding proteins bind to the DNA double helix by inserting an α -helix into the major groove (Pabo and Sauer 1992 *Annu. Rev. Biochem.* 61:1053-1095; Harrison 1991 *Nature (London)* 353: 715-719; and Klug 1993 *Gene* 135:83-92). Sequence specificity results from the geometrical and chemical complementarity between the amino acid side chains of the α -helix and the accessible groups exposed on the edges of base-pairs. In addition to this direct reading of the DNA sequence interactions with the DNA backbone stabilize the complex and are sensitive to the conformation of the nucleic acid, which in turn depends on the base sequence (Dickerson and Drew 1981 *J. Mol. Biol.* 149:761-786)

Zfs have become the DBD of choice in efforts to engineer custom-made transcription factors. A Zf is an independently folded zinc-containing mini-domain, the structure of which is well known in the art and defined in, for example, Miller et al., (1985) *EMBO J.* 4:1609; Berg (1988) *Proceedings of the National Academy of Sciences (USA)* 85:99; Lee et al., (1989) *Science* 245:635 and Klug, (1993) *Gene* 135:83. The crystal structures of Zf DNA complexes show a semi-conserved pattern of interactions, in which typically 3 amino acids from the α -helix of the Zf contact 3 adjacent base pairs (bp) or a "subsite" in the DNA (Pavletich et al., (1991)

Science 252:809; Fairall et al., (1993) Nature 366:483; and Pavletich et al., (1993) Science 261:1701). Thus, the crystal structure of Zif268 suggested that Zf DBDs might function in a modular manner with a one-to-one interaction between a Zf and a 3 bp "subsite" in the DNA sequence. In naturally occurring transcription factors, multiple Zfs are typically linked together in a tandem array to achieve sequence-specific recognition of a contiguous DNA sequence (Klug, (1993) Gene 135:83).

Multiple studies have shown that it is possible to artificially engineer the DNA binding characteristics of individual Zfs by randomizing the amino acids at the α -helical positions involved in DNA binding and using selection methodologies such as phage display to identify desired variants capable of binding to DNA target sites of interest (Rebar et al., (1994) Science 263:671; Choo et al., (1994) Proceedings of the National Academy of Sciences (USA) 91:11163; Jamieson et al., (1994) Biochemistry 33:5689; Wu et al., (1995) Proceedings of the National Academy of Sciences (USA) 92: 344). Similarly, there are numerous patents and patent applications relating to methods of producing and using artificially engineered zinc finger proteins, see for Example, Published U.S. patent application s2002/0160940 A1 and 2002/0164575 A1, U.S. Patent Nos. 6,511,808, 6,013,453, 6,007,988, 6,503,717, 6,453,242, 9, U.S. Patent No. 6,492,117, and International publications WO 02099084A2, WO 02089498, WO 02057308 A2, WO 0153480 A and WO 0027878 A1.

Furthermore, by fusing such recombinant Zf DBDs to regulatory or effector domains, it has been possible to artificially regulate expression of transfected reporter genes in cultured cells. For example, Beerli et al., (Beerli et al., (1998) Proceedings of the National Academy of Sciences (USA) 95:14628) reported construction of a chimeric 6 finger Zf protein fused to either a KRAB, ERD, or SID transcriptional repressor domain, or the VP16 or VP64 transcriptional activation domain. This chimeric Zf protein was designed to recognize an 18 bp target site in the 5' untranslated region of the human erbB-2 gene. Using this construct, the authors were able to either activate or repress a transiently expressed reporter luciferase construct linked to the erbB-2 promoter. Although these proteins were designed to recognize an 18 bp DNA sequence, recent evidence demonstrates that they specify only some of the 18 bases in their target site (see below and Segal et al. (2003) Biochemistry 42 (7) 2317-2148).

Further studies have demonstrated that such recombinant Zf transcription factors can also be used to regulate expression of endogenous genes in their native chromosomal context (see for

Example, Reik et al., (2002) *Current Opinions in Genetics & Development* 12:233, and Published U.S. patent application 2002/0160940 A1). Clinically relevant human genes that have been successfully regulated in this way include MDR1, erythropoietin, erbB-2 and erbB-3, VEGF, and PPARgamma. In the case of VEGF (Liu et al., (2001) *Journal of Biological Chemistry* 276:11323), proportional up-regulation by the designed transcription factor of all 3 distinct splice isoforms generated by this locus was observed, illuminating the utility of endogenous gene control in therapeutic settings (proper isoform ratio is essential for the proangiogenic function of VEGF). Furthermore, Rebar et al., (*Nature Medicine* 8: 1427-1432 (2002) showed that Zf transcription factors designed to bind to the VEGF-A gene induced expression of VEGF-A *in vivo* leading to a stimulation of angiogenesis and an acceleration of experimental wound healing. In the case of PPARgamma, use of a transcriptional repressor designed to downregulate the expression of 2 PPARgamma isoforms allowed "mutation-free reverse genetics" analysis that illuminated a unique role for the PPARgamma2 isoform in adipogenesis (Ren et al., (2002) *Genes & Development* 16:27).

In order to use recombinant Zfs to specifically target a gene of interest within the genome, the target site sequence recognized should be sufficiently long that statistically it occurs only once in the genome. In the case of the human genome, a multi-finger Zf protein recognizing a stretch of about 16 bp or more should be generated for this to be achieved (Liu et al., (1997) *Proceedings of the National Academy of Sciences (USA)* 94:5525). Statistically, assuming random base distribution, a unique 16 bp sequence will occur only once in 4.3×10^9 bp, thus a 16 bp sequence should be sufficient to specify a unique address within the approximately 3.5×10^9 bp that make up the human genome (Liu et al., (1997) *Proceedings of the National Academy of Sciences (USA)* 94:5525). Similarly, an 18 bp address specified by a 6 finger protein, would enable sequence specific targeting within 6.8×10^{10} bp of DNA. Such a 6-finger protein would thus be able to uniquely specify any locus within all currently known genomes, and thus could be used to artificially regulate the expression of only an intended target gene and not other unintended genes or sequences.

Various strategies have been described for creating multi-finger proteins capable of binding such extend DNA sequences. The majority of such strategies have involved linking together tandem arrays of engineered zinc fingers derived from naturally occurring proteins that contain 3 zinc fingers in their DNA binding domains, such as Zif268 and Sp1, see for example,

U.S. patent application 2002/0160940 A1, WO 02099084A2, and WO 0153480 A1. Zinc finger units from these proteins are typically linked together using the canonical 5 amino acid zinc finger linker sequence, TGEKP, to generate proteins composed of 3, 4, 5, 6 and 9 fingers (U.S. Patent No. 6,140,466). However, biochemical characterization of these synthetic multi-finger proteins has revealed an apparent energetic barrier to the simultaneous binding of more than 3 fingers to a target DNA sequence. For proteins composed of fingers derived from Zif268 or Sp1 connected by standard TGEKP linkers, the binding energy increases dramatically (by approximately 3 orders of magnitude) as one progresses from a 2-finger domain to a 3-finger domain. However, it has been found that adding more fingers to a 3-finger domain does not yield the expected large increase in binding affinity. This result has been observed with 4-, 6-, and 9-finger proteins. The precise reason for this barrier is not well understood although it is known that using longer, non-TGEKP linkers between selected finger units can restore some, but not all, of the expected affinity for a 6 finger protein (see, for example, WO 0153480 A1). These results suggest that no more than 3 fingers linked by TGEKP linkers can bind simultaneously to DNA. An important implication of this finding is that synthetic TGEKP-linked proteins with 4 or more Zfs do not specify all of the DNA bases in their intended target DNA sequence, and therefore may not be suitable for applications where it is important that only the specific target gene of interest is bound and regulated.

The findings of Segal et al. (Biochemistry 42 (7) 2317-2148 (2003)) provide further evidence that currently available methods for producing multi-finger proteins will not be suitable for the production of artificial transcription factors capable of regulating specific genes in humans. As described above, statistically an 18 bp address specified by a 6 finger protein should be able to uniquely specify any locus within the human genome. However, Segal et al. found that when they linked together two 3-finger Zf proteins, each of which had perfect specificity for its 9 bp target site, the specificity of the 6 finger protein was significantly less than predicted. Furthermore, Segal et al. calculated that the consequence of the reduction in specificity that they observed was that instead of being able to uniquely specify a single 18 bp sequence, the 6-finger proteins they produced would bind to roughly 66 different sites in the human genome.

Segal et al. suggested several possible explanations for this effect. These included a) the possibility that the increased number of contacts in the 6-finger protein elevated the binding energy to a point where individual residue:base mismatches were insufficient to prevent binding,

b) the possibility that having so many contacts to one strand of the DNA was sufficient to "pull" the protein toward that strand and mis-orient some of the fingers, and c) that the DNA-contacting residues of the longer protein fail to align properly with the DNA bases, thus limiting the number of fingers that can engage simultaneously with the DNA target site. This latter idea is consistent with the observation (noted above) that there appears to be an energetic barrier to simultaneously engaging more than 3 zinc fingers to a DNA site using existing methodologies and finger scaffolds for design.

Another problem with current methods of producing multi-finger proteins is that of binding affinity. Previous studies have shown that only proteins that bind their target with a dissociation constant in the nanomolar to picomolar range or better are able to effectively regulate gene expression. However, it is also likely to be true that if binding affinity is too great gene expression can not be effectively regulated either. If the dissociation constant of a protein is too low (i.e. affinity is too high), then at the physiologic levels of protein expression a zinc finger protein will likely occupy DNA sites other than its intended target sequence. In addition, if affinity is too high (e.g. a dissociation constant in the femtomolar range), it is possible that a "kinetic trapping" effect may occur where the zinc finger protein becomes "stuck" to unintended binding sites in the genome (see Kim and Pabo, PNAS (1998) 95(6):2812-7, for example). Thus, in order to be useful in regulating gene expression, ideally an artificial Zf protein should bind to a unique target sequence in the human genome with a dissociation constant in the picomolar to nanomolar range. Predictions based on the chelate effect suggest that if 6 Zfs derived from a 3-finger protein such as zif268 or Sp-1 were strung together in such a way that all fingers simultaneously bound to the DNA, the dissociation constant of the resultant protein would be on the order of 10^{-18} to 10^{-21} molar. Thus another problem with engineering multi-finger proteins using fingers from naturally occurring 3-finger proteins is that even if one could find a means to permit more than three zinc fingers to simultaneously engage their target DNA site, the binding affinity of the resulting protein would likely be too high to be useful.

Choo and colleagues (Moore et al., PNAS (2001) Feb 13;98(4):1437-41) have described a method for producing six finger proteins in which three two-finger units (derived from two fingers of Zif268) are connected together using a TGGEKP linker. The use of a non-TGEKP linker disrupts the ability of any subset of fingers within one of these proteins from binding to the DNA as a three-finger unit. The overall affinity of the resulting six-finger proteins is in a

physiologically useful range. However, it remains unclear whether all six fingers in these proteins are simultaneously engaging the DNA site and also the overall specificity of these proteins remains unclear.

The Neuron Restrictive Silencer Factor (hereafter referred to as NRSF), which is also known as the RE-1 Silencing Transcription Factor or REST, is a naturally occurring zinc finger protein that binds to a 21 bp DNA sequence called the Neuron Restrictive Silencer Element (hereafter referred to as NRSE). The NRSF protein, first identified by Chong et al., (Cell 80 (6) 949-957 (1995)) and Schoenherr et al, (Science 267 (5202) 1360-1363 (1995)), is described in U.S. Patents Nos. 5,935,811 and 6,270,990.

This sequence is found in many genes that encode proteins required for neuronal function such as the type II sodium channel gene, the SCG10 gene. A list of many the genes that contain NRSE sequences can be found in Schoenherr et al., 1996 (PNAS 93; p 9881-9886). The NRSF protein is predominantly expressed in non-neuronal cells. It functions as a master regulator of neuronal gene expression by repressing the expression of its target genes in non-neuronal cells. In addition to its DNA binding domain consisting of 8 tandem Cys₂His₂ zinc fingers, the NRSF protein also comprises 2 repression domains, one located at each end of the protein. By differential utilization of these repression domains, NRSF mediates both active repression and long-term silencing of its target genes. Thus, NRSF provides a naturally occurring example of a multi-finger protein that can recognize an extended 21 base pair binding site with high specificity.

To date, very little biochemical or genetic information, and no structural information, exists about the interaction of the 8-finger NRSF DNA binding domain with the NRSE. One study, in which entire fingers were inactivated by substituting an arginine for one of the conserved cysteines, revealed that neutralization of NRSF finger 7, or of a combination of fingers 6 and 8, leads to diminished DNA binding by NRSF. Another study provided evidence suggesting that a splice form of NRSF that contains fingers 3 through 5 can bind near the 3' end of the NRSE but no detailed mapping of finger-DNA interactions was performed.

OBJECTS AND/OR SUMMARY OF THE INVENTION

The present invention provides methods for selecting multi-finger Zf proteins having 4 - 8 Zfs that bind selectively to DNA target sequences of 12 to 24 bp. Currently known methods of

designing or selecting non-naturally occurring Zf proteins typically start with a natural Zf protein as a source of "scaffold" residues. The process of design or selection serves to alter the amino acid composition of the Zfs so as to confer the desired DNA binding specificity. The majority of currently used methods for designing or selecting Zf proteins having 4 or more zinc fingers utilize 3-finger Zf proteins, such as zif268, as the "scaffold" and produce proteins having more than 3 fingers by linking the desired number of Zfs together using the well conserved canonical Zf linker sequence, TGEKP, or other synthetic linkers (see above). However, as noted above, biochemical analysis to date suggests that no more than 3 fingers in these proteins can bind simultaneously to DNA. This means that such proteins do not specify all of the DNA bases in their intended target DNA sequence, and therefore may not be suitable for applications where it is important that only the specific target gene of interest is bound and regulated. Furthermore, studies show that the binding affinity of the individual fingers in naturally occurring 3-finger proteins is such, that if 6 fingers were joined together and simultaneously bound to their DNA target, the binding affinity of the resultant protein would be too high to allow productive regulation of gene expression.

It is an object of the present invention to create synthetic proteins that bind to extended DNA target sites with high affinity and specificity, not by linking together multiple fingers from three-finger proteins (e.g., zif268) as has been done previously but by re-engineering the DNA binding specificity of a naturally occurring multi-finger Zf protein that binds to an extended DNA target sequence.

The methods of the present invention overcome these problems in the art by using as a scaffold a protein that has 8 Zfs in its DNA-binding domain, namely NRSF. Thus, the present invention is unique in exploiting the zinc fingers and "linkers" from a Zf protein that has naturally evolved (and is therefore presumably optimized) to bind to an extended DNA target sequence with an affinity in the physiologically useful range (e.g. with a dissociation constant in the nanomolar to picomolar range.).

To date there have been very few studies on the interaction of the NRSF Zfs with its NRSE target sequence. Similarly, there have been no attempts to alter the DNA binding specificity of the NRSF protein. Surprisingly, it now has been shown that NRSF actually binds simultaneously to a span of at least 20 DNA bases, suggesting that NRSF is uniquely able to bind to extend DNA sequences with high specificity. It has also been surprisingly found that binding

requires only zinc fingers 3-8 of the NRSF protein. Furthermore, the specific nucleotide contacts made by each of zinc finger 3-8 in NRSF have now been modeled and specific DNA contacts made by fingers 4 and 5 have been defined. This critical new information on the binding of NRSF to its target sequence has enabled the development of the novel Zf engineering methods of the present invention.

Accordingly, in one aspect, the present invention provides methods of selecting synthetic zinc-finger polypeptides that bind specifically to DNA target sequences of interest, where the NRSF protein is used as the scaffold sequence.

In another aspect, the present invention provides preferred libraries and screening methods to be used in the production of such NRSF-based synthetic Zf proteins.

In another aspect, the invention is directed to methods of selecting appropriate target sequences within a gene of interest. The invention provides criteria and methods for selecting optimum subsequence(s) from a target gene for targeting by a NRSF-based zinc finger protein.

BRIEF DESCRIPTION OF THE DRAWINGS

In the following Detailed Description and Examples reference will be made to the accompanying drawings, incorporated herein by reference.

Figure 1 provides the amino acid sequence of the human NRSF protein (SEQ ID NO. 1).

Figure 2 provides the sequence of the human the NRSF cDNA (SEQ ID NO. 2).

Figure 3 provides the amino acid sequence of the mouse NRSF protein (SEQ ID NO. 3)

Figure 4 provides the sequence of the mouse NRSF cDNA (SEQ ID NO. 4)

Figure 5 provides the amino acid sequence of the rat NRSF protein (SEQ ID NO. 5)

Figure 6 provides the sequence of the rat NRSF cDNA (SEQ ID NO. 6)

Figure 7 provides the amino acid sequence of the *Xenopus laevis* NRSF protein (SEQ ID NO. 7)

Figure 8 provides the sequence of the *Xenopus laevis* NRSF cDNA (SEQ ID NO. 8)

Figure 9 shows the amino acid sequences of each of the zinc fingers and inter-finger linkers in Zif268 and human NRSF (taken from SEQ ID No. 1). The arrows and barrel above the sequences indicate positions of β -sheet and α -helix respectively. Conserved cysteines and histidines are shown in red and purple, respectively. Recognition helix sequences (including the

-1 residue preceding the helix start) are underlined. Linker sequences are shown in blue. The unusual tyrosine in the NRSF fingers is shown in green.

Figure 10 provides a schematic representation of the bacterial two-hybrid method, showing an *E. coli* cell bearing a single copy reporter construct with a "target sequence" positioned upstream of a weak promoter. A zinc finger protein fused to a Gal11P fragment and RNA polymerase (RNAP) α -subunit hybrid proteins containing a fragment of the Gal4 protein are expressed in the cell. Recruitment of RNAP by the zinc finger protein leads to activation of the reporter gene(s) and is mediated by the interaction between the Gal11P and Gal4.

Figure 11 illustrates binding of NRSF1-8 and NRSF3-8 domains to various NRSE sites in the bacterial two-hybrid system. As described in the Examples, GP-NRSF1-8 and GP-NRSF3-8 proteins were tested in "B2H reporter strains" harboring the depicted NRSE sequence. Mutated bases in the NRSE sites are shown in red. Fold-activation indicates the extent of transcriptional activation of the *lacZ* reporter gene in the strains.

Figure 12 shows the predicted model for the interaction of NRSF fingers 3-8 with the consensus NRSE. Recognition helix residues -1, 2, 3, and 6 from fingers 3-8 are shown with postulated contacts (arrows) to the consensus NRSE sequence. Strongly conserved positions in the NRSE are shown in uppercase whereas less well conserved positions are in lowercase.

Figure 13 shows electrophoretic mobility shift assay (EMSA) data showing binding of NRSF3-8 and NRSF 1-8 to the NRSE.

Figure 14 shows sequences of re-engineered NRSF-based variants. Sequences of residues selected in recognition helices of re-engineered fingers are shown. Finger 4 variants are shown in Figure 14A and Finger 5 variants are shown in Figure 14B. The double mutant NRSE targeted is shown above the finger sequences. Mutations are shown in red.

Figure 15 provides data showing that re-engineered NRSF-based variants bind specifically to their target mutant NRSE sequence. Finger 4 (Figure 15A) and Finger 5 (Figure 15B) NRSF-based variants were introduced into B2H reporter strains harboring the consensus NRSE, the appropriate double mutant target NRSE, or a point mutant NRSE. The ability of each NRSF variant to activate transcription in the indicated reporter strain is expressed as fold-activation of the promoter.

Figure 16 depicts an overview of the CSPO Multi-Finger Optimization Strategy. Proteins with greater numbers of fingers can also be optimized by performing additional primary selections.

Figure 17 depicts a schematic representation of experiments to assess the activity of selected NRSF-based variants (described in Example 9) in mammalian cells. As described in the text, for each of the target DNA sequences used, several selected NRSF-based variants can be used to construct stable cell lines that express these proteins. For each selected NRSF-based variant, multiple stable cell lines will be identified. RNA extracted from each of these cell lines is hybridized to an Affymetrix U133A GeneChip.

Figure 18 depicts a schematic of microarray experiments used to provide insight into the functional specificity of a three-finger protein in a mammalian cell, as in Example 9. The data shown comes from a single microarray experiment using Affymetrix U133A chips which was performed to assess the global effects of a three-finger protein (VZ-573) fused to a transcriptional activator domain. Three sets of 30 genes each were selected: the 30 unique genes with the greatest fold activation ("activated genes"), the 30 genes whose expression levels were apparently unaffected ("unaffected genes"), and the 30 genes with the greatest fold-repression ("repressed genes"). Genomic sequences flanking the likely transcriptional start sites for each gene were obtained and searched on both strands for matches or near-matches (off by one base, at either of two positions judged most likely to be degenerate for this protein). The average number of matches per gene within 2500 bases of the transcriptional start site (shown in spans of 500 base pairs) is shown for the three different set of genes.

Figures 19 to 32 show the amino acid sequences of selected NRSF-based proteins F4v1 F4v4, F4v5, F4v6, F4v7, F4v8, F5v1, F5v2, F5v3, F5v4, F5v5, F5v6, F5v7 and F5v8, respectively, as described in Example 5.

Figure 33 shows sequences of a) wild-type and b) re-engineered variants of NRSF finger 6 with the relevant portions of their associated consensus or mutant NRSE sequences, as described in Example 7. Arrows indicate contacts consistent with interactions observed in previously described zinc finger-DNA interfaces.

Figure 34 shows sequences of a) wild-type and b) re-engineered variants of NRSF finger 8 with the relevant portions of their associated consensus or mutant NRSE sequences, as

described in Example 7. Arrows indicate contacts consistent with interactions observed in previously described zinc finger-DNA interfaces.

Figure 35 b) summarizes NSRF-NRSE interactions as determined from the NRSF finger re-engineering data presented herein, in comparison with predicted interactions (part a).

DETAILED DESCRIPTION

I. Introduction

The present invention provides engineered multi-finger Zf polypeptides capable of binding to an extended target DNA sequence within a gene of interest, and methods of selection thereof. The scaffold Zf protein that is used as the starting point from which to engineer and select these Zf polypeptides is the naturally occurring transcription factor NRSF which has a DNA binding domain comprising 8 Zfs. Using the methods of the present invention, all of the Zfs of NRSF can be engineered and selected for binding to a given sequence of interest, enabling the construction of an engineered Zf protein capable of binding specifically to a DNA target sequence spanning up to 21 bp. The present invention also provides methods for selection of suitable target sequences within genes of interest. Further details of the methods of the present invention are provided below.

II. Definitions

As used herein, the following terms have the meanings ascribed to them unless specified otherwise.

In this disclosure, "comprises," "comprising," "containing" and "having" and the like can have the meaning ascribed to them in U.S. Patent law and can mean "includes," "including," and the like; "consisting essentially of" or "consists essentially" likewise has the meaning ascribed in U.S. Patent law and the term is open-ended, allowing for the presence of more than that which is recited so long as basic or novel characteristics of that which is recited is not changed by the presence of more than that which is recited, but excludes prior art embodiments.

The term "zinc finger" or "Zf" refers to a polypeptide having DNA binding domains that are stabilized by zinc. The individual DNA binding domains are typically referred to as "fingers." A Zf protein has at least one finger, preferably 2 fingers, 3 fingers, or 6 fingers. A Zf protein having 2 or more Zfs is referred to as a "multi-finger" or "multi-Zf" protein. Each finger typically comprises an approximately 30 amino acid, zinc-chelating, DNA-binding domain. An exemplary motif characterizing one class of these proteins is -Cys-(X) (2-4)-Cys-(X) (12)-His-

(X) (3-5)-His (SEQ ID NO:9), where X is any amino acid, which is known as the "C(2)H(2)class." Studies have demonstrated that a single Zf of this class consists of an alpha helix containing the 2 invariant histidine residues co-ordinated with zinc along with the 2 cysteine residues of a single beta turn (see, e.g., Berg and Shi, Science 271:1081-1085 (1996)).

The present invention relates to the "Neuron Restrictive Silencer Factor" or "NRSF" protein, which is also known as the "RE-1 silencing transcription factor" or "REST". Unless otherwise specified, the name "NRSF" as used herein refers to NRSF proteins, or nucleic acids encoding NRSF proteins, from any animal species. Thus the name "NRSF" encompasses, for example, human, rat and mouse NRSF proteins. Furthermore, the name "NRSF" encompasses all splice variants of the NRSF protein, several of which are known (see for example, Palm et al., J. Neurosci 15: 1280-96 (1998) and Palm et al., Brain Res 8: 72 (1999)). In addition, the name "NRSF" includes homologues and fragments of the NRSF protein.

The term "homologues" as used herein refers to nucleic acid or protein sequences sharing a certain amount of sequence homology or identity with a wild-type NRSF protein. The terms "homology" and "identity" with respect to a nucleotide or amino acid sequence indicate a quantitative measure of homology between two sequences. The percent sequence homology can be calculated as $(N_{ref} - N_{dif}) * 100 / N_{ref}$, wherein N_{dif} is the total number of non-identical residues in the two sequences when aligned and wherein N_{ref} is the number of residues in one of the sequences. Hence, the DNA sequence AGTCAGTC will have a sequence identity of 75% with the sequence AATCAATC ($N_{ref} = 8$; $N_{dif} = 2$).

Alternatively or additionally, "homology" or "identity" with respect to sequences can refer to the number of positions with identical nucleotides or amino acids divided by the number of nucleotides or amino acids in the shorter of the two sequences wherein alignment of the two sequences can be determined in accordance with the Wilbur and Lipman algorithm (Wilbur and Lipman, 1983 PNAS USA 80:726, incorporated herein by reference), for instance, using a window size of 20 nucleotides, a word length of 4 nucleotides, and a gap penalty of 4, and computer-assisted analysis and interpretation of the sequence data including alignment can be conveniently performed using commercially available programs (e.g., Intelligenetics™ Suite, Intelligenetics Inc. CA).. When RNA sequences are said to be similar, or have a degree of sequence identity or homology with DNA sequences, thymidine (T) in the DNA sequence is considered equal to uracil (U) in the RNA sequence. Thus, RNA sequences are within the scope

of the invention and can be derived from DNA sequences, by thymidine (T) in the DNA sequence being considered equal to uracil (U) in RNA sequences.

Advantageously, sequence identity or homology such as amino acid sequence identity or homology can be determined using the BlastP program (Altschul *et al.*, Nucl. Acids Res. 25, 3389-3402, incorporated herein by reference) and available at NCBI, as well as the same or other programs available via the Internet at sites thereon such as the NCBI site.

The following documents (each incorporated herein by reference) provide algorithms for comparing the relative identity or homology of sequences such as amino acid residues of two proteins, and additionally or alternatively with respect to the foregoing, the teachings in these references can be used for determining percent homology or identity: Needleman SB and Wunsch CD, "A general method applicable to the search for similarities in the amino acid sequences of two proteins," J. Mol. Biol. 48:444-453 (1970); Smith TF and Waterman MS, "Comparison of Bio-sequences," Advances in Applied Mathematics 2:482-489 (1981); Smith TF, Waterman MS and Sadler JR, "Statistical characterization of nucleic acid sequence functional domains," Nucleic Acids Res., 11:2205-2220 (1983); Feng DF and Dolittle RF, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," J. of Molec. Evol., 25:351-360 (1987); Higgins DG and Sharp PM, "Fast and sensitive multiple sequence alignment on a microcomputer," CABIOS, 5: 151-153 (1989); Thompson JD, Higgins DG and Gibson TJ, "ClusterW: improving the sensitivity of progressive multiple sequence alignment through sequence weighing, positions-specific gap penalties and weight matrix choice," Nucleic Acid Res., 22:4673-480 (1994); and, Devereux J, Haeberlie P and Smithies O, "A comprehensive set of sequence analysis program for the VAX," Nucl. Acids Res., 12: 387-395 (1984). And, without undue experimentation, the skilled artisan can consult with many other programs or references for determining percent homology.

In one embodiment, the NRSF homologues of the present invention share at least 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% sequence homology the wild-type NRSF protein of the corresponding species. Preferably the NRSF homologues share at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% homology. More preferably the NRSF homologues share at least 90%, 95%, 96%, 97%, 98%, or 99% of their sequence with that of a wild-type NRSF protein. More preferably still, the NRSF homologues share 95%, 96%, 97%, 98%, or 99% of their sequence with a wild-type NRSF protein.

The homology to a wild-type NRSF protein need not span the entire length of the NRSF protein. A key feature of the present invention is that only the zinc finger DNA binding domain of NRSF need be used in the construction of NRSF-based zinc finger proteins. Therefore, the present invention only requires that the above degrees of homology relate to the amino acid or nucleotide sequence of the zinc finger DNA binding domain of a wild-type NRSF protein. The only requirement is that the homologous sequences still encode zinc finger proteins.

A "functional" homologue or fragment of an NRSF protein, polypeptide or nucleic acid is a protein, polypeptide or nucleic acid whose sequence is not identical to the full-length wild-type NRSF protein, polypeptide or nucleic acid, but yet retains some of the same functions as the full-length NRSF protein, polypeptide or nucleic acid. In particular, in the methods of the present invention, a "functional homologue" is one that encodes a protein that conforms to a zinc finger consensus sequence, and is capable of binding to DNA. A functional fragment can possess more, fewer, or the same number of residues as the corresponding native molecule, and/or can contain one or more amino acid or nucleotide substitutions. Methods for determining the function of a nucleic acid (e.g., coding function, ability to hybridize to another nucleic acid) are well-known in the art. Similarly, methods for determining protein function are well-known. For example, the DNA-binding function of a polypeptide can be determined, for example, by filter-binding, electrophoretic mobility-shift, or immunoprecipitation assays. See Ausubel et al., *supra*. The ability of a protein to interact with another protein can be determined, for example, by co-immunoprecipitation, two-hybrid assays or complementation, both genetic and biochemical. See, for example, Fields et al. (1989) *Nature* 340:245-246; U.S. Pat. No. 5,585,245 and PCT WO 98/44350.

The term "linker" or "inter-finger linker" as used herein refers to a stretch of amino acids located between 2 Zfs in a given protein or polypeptide. In certain embodiments of the present invention, selected zinc finger proteins are covalently linked together using such amino acid linkers. In other embodiments, selected zinc finger proteins are non-covalently linked by the process of "multimerization" or "dimerization". As used herein "multimerization" refers to the non-covalent linkage of more than two individual proteins or polypeptides, while "dimerization" refers to the non-covalent linkage of only two individual proteins or polypeptides. The individual proteins that are linked together may be identical to each other, in which case the proteins are said to "homo-multimerize" or "homo-dimerize", or they may be different, in which case the

proteins are said to "hetero-multimerize" or hetero-dimerize". The protein complexes produced by such non-covalent linkages are referred to as "multimers" or "dimers," respectively. The production of such a zinc finger multimer or dimer may be performed by fusion of a "multimerization domain" or "dimerization domain" to a selected zinc finger protein. Such domains are amino acid sequences that when present in a polypeptide cause that polypeptide to multimerize or dimerize.

The DNA sequence to which a Zf protein binds is referred to as a "target site" or "target sequence." Each Zf within a Zf protein binds from about 2 to about 5 base pairs within the target site, preferably 3 or 4 base pairs (the "subsite"). Accordingly, a "subsite" is a subsequence of the target site, and corresponds to a portion of the target site bound by a single finger. A single Zf preferably recognizes a 3 or 4 bp subsite. An "extended target site" or "extended target sequence" as used herein refers to a sequence of DNA for which more than 3 zinc fingers are required for binding, thus it refers to a sequence 10 bp or greater in length. The methods of the present invention can be used to generate Zf proteins having up to 8 zinc fingers that bind to an extended target sequence of up to around 20-24 bp.

The target site that is bound by the naturally occurring transcription factor NRSF is known as a Neuron Restrictive Silencer Element or "NRSE". The exact nucleotide sequence of the NRSE sequence that NRSF binds varies between the group of 50 or so genes that are regulated by NRSF. However, a consensus NRSE sequence has been derived from nineteen different experimentally conserved NRSF binding sites: 3'CcgcGAcAGGcaCCACGACTt5' (SEQ ID NO. 10). Thirteen positions in this consensus NRSE are strongly conserved (uppercase letters) and 8 are more weakly conserved (lowercase letters). None of the nineteen sequences used to derive the consensus differ by more than 6 bases from the consensus. Examination of experimentally confirmed NRSE sequences reveals that very few differ from the consensus by more than 3 bases (and these differences typically occur in the more weakly conserved positions). As used herein the term "NRSE" refers to any sequence fitting this consensus sequence (SEQ ID NO. 10).

The term "scaffold" as used herein refers to a naturally occurring Zf protein, or a portion thereof, that is used as the starting point from which to engineer a new Zf protein by altering the amino acid sequence of the scaffold protein. The methods of the present invention use the naturally occurring transcription factor NRSF, or a Zf-containing portion thereof, as a scaffold.

The proteins or polypeptides that are generated by alteration of the amino acid sequence of the NRSF scaffold are referred to as "NRSF-based" or "NRSF-derived." The methods of the present invention involve generating new NRSF-based Zf proteins that bind to a chosen target site that differs from the NRSE consensus sequence. The terms "designed" "engineered" "synthetic" "artificial" and "non-naturally occurring" as used herein refer to Zf proteins that have been generated or selected to bind to a given target sequence that is not the sequence bound by the scaffold protein, and which differ in amino acid sequence from the scaffold protein.

The term "library" as used herein refers to a population of nucleic acid sequences that encode Zf polypeptides. Such "libraries" are used in the present invention to screen for and identify Zf polypeptides having desired characteristics from large and complex pool of Zf polypeptides. Such libraries can be created in cell free systems or within eukaryotic cells, prokaryotic cells or viral particles. The term "primary library" refers to a library that has not been enriched for nucleic acids encoding Zf polypeptides with particular characteristics. The term "secondary library" refers to a library that is enriched for nucleic acids encoding Zf polypeptides with particular characteristics.

The term "randomized" or "randomize" refers to a pool of Zf molecules, or the generation of a pool of Zf molecules, in which one of a multitude of possible amino acids is represented at one or more given "variable" amino acid positions.

" K_D " refers to the dissociation constant for binding of one molecule to another molecule, i.e., the concentration of a molecule (such as a Zf protein), that gives half maximal binding to its binding partner (such as a DNA target sequence) under a given set of conditions. The K_D provides a measure of the strength of the interaction between 2 molecules, or the "affinity" of the interaction between 2 molecules. 2 molecules that bind strongly to each other have a "high affinity" for each other, while molecules that bind weakly to each other have a "low affinity" for each other.

The term "recombinant" when used herein with reference to portions of a nucleic acid or protein, indicates that the nucleic acid comprises 2 or more sub-sequences that are not found in the same relationship to each other in nature. For instance, a nucleic acid that is recombinantly produced typically has 2 or more sequences from distinct genes or non-adjacent regions of the same gene, synthetically arranged to make a new nucleic acid sequence encoding a new protein, for example, a DBD from one source and a regulatory or effector region from another source, or

a Zf from the native Zif268 protein and a Zf selected from a library. The term "recombination" as used herein, refers to the process of producing a recombinant protein or nucleic acid by standard techniques known to those skilled in the art, and described in, for example, as Sambrook et al., *Molecular Cloning; A Laboratory Manual* 2d ed. (1989).

"Nucleotide" refers to a base-sugarphosphate compound. Nucleotides are the monomeric subunits of both types of nucleic acid molecules, RNA and DNA. Nucleotide refers to ribonucleoside triphosphates, rATP, rGTP, rUTP and rCTP, and deoxyribonucleoside triphosphates, such as dATP, dGTP, dTTP, and dCTP.

"Base" refers to the nitrogen-containing base of a nucleotide, for example adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U). "Base pair" or "bp" refers to the partnership of bases within the DNA double helix, whereby typically an A on one strand of the double helix is paired with a T on the other strand and a C on one strand of the double helix is paired with a G on the other strand.

"Nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring, which have similar binding properties as the reference nucleic acid, and which are metabolized in a manner similar to the reference nucleotides. Examples of such analogs include, without limitation, phosphorothioates, phosphoramidates, methyl phosphonates, chiral-methyl phosphonates, 2-O-methyl ribonucleotides, peptide-nucleic acids (PNAs). Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (e.g., degenerate codon substitutions) and complementary sequences, as well as the sequence explicitly indicated. The term nucleic acid is used interchangeably with gene, cDNA and nucleotide. The nucleotide sequences are displayed herein in the conventional 5' to 3' orientation.

The terms "polypeptide," "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an analog or mimetic of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers. Polypeptides can be modified, e.g., by the addition of carbohydrate residues to form glycoproteins. The terms "polypeptide," "peptide" and "protein" include glycoproteins, as well as non-glycoproteins. The polypeptide

sequences are displayed herein in the conventional N-terminal to C-terminal orientation.

The term "amino acid" refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, e.g., hydroxyproline, carboxyglutamate, and O-phosphoserine. Amino acid analogs refers to compounds that have the same basic chemical structure as a naturally occurring amino acid, i.e., a carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, e.g., homoserine, norleucine, methionine sulfoxide, methionine, and methyl sulfonium. Such analogs have modified R groups (e.g., norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. Amino acid mimetics refers to chemical compounds that have a structure that is different from the general chemical structure of an amino acid, but that functions in a manner similar to a naturally occurring amino acid. The terms "amino acid residue" or "residue" refer to a specific amino acid position within a polypeptide or protein.

Degenerate codon substitutions or "doping strategies" may be achieved by generating sequences in which any position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzner et al., *Nucleic Acid Res.* 19:5081 (1991); Ohtsuka et al., *J. Biol. Chem.* 260:2605-2608 (1985); Rossolini et al., *Mol. Cell. Probes* 8:91-98 (1994)). Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given protein. For instance, the codons GCA, GCC, GCG and GCU all encode the amino acid alanine. Thus, at every position where an alanine is specified by a codon in an amino acid herein, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of conservatively modified variations. Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation of the nucleic acid. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine, and TGG, which is ordinarily the only codon for tryptophan) can be modified to yield a functionally identical molecule. Accordingly, each silent variation of a nucleic acid which encodes a polypeptide is implicit in each described sequence.

"Specific" or "specific-binding" as used herein, refers to the interaction between a protein and a nucleic acid wherein the protein recognizes and interacts with a defined nucleotide

sequence or sequences, as opposed to a “non-specific” interaction wherein the protein does not require a defined nucleotide sequence to associate with the nucleic acid molecule (for example, a protein that interacts with the phosphate-sugar backbone of the DNA but not the bases of the nucleotides). The strength of the association between the protein and the nucleic acid molecule can vary significantly between different “binding complexes.” A “binding complex,” as used herein, comprises an association between a sequence of interest, target site or subsite and a Zf binding domain. “Binding complexes” can comprise both weakly-bound Zf proteins and nucleic acids and strongly-bound Zf proteins and nucleic acids. The strength or “affinity” of the association of a Zf with an intended or specified sequence of interest, target site or subsite is expressed in terms of the K_D , as defined above.

“Conditions sufficient to form binding complexes” refers to the physical parameters selected for a binding reaction or “incubation” between a nucleic acid and a protein sample that potentially contains an unknown nucleic acid-binding protein, such as, buffer ionic strength, buffer pH, temperature, incubation time, and the concentrations of nucleic acid and protein, where such physical parameters allow nucleic acids to bind to proteins. Such conditions can be “low-stringency conditions”, which are conducive to the formation of “binding complexes” comprising both weakly- and strongly-bound proteins and nucleic acids or “high-stringency conditions”, which are conducive to the formation of “high affinity binding complexes” comprising only strongly-bound proteins and nucleic acids. Low-stringency conditions typically comprise high salt concentration and a temperature ranging between 37C and 47C. When DNA-protein “binding reactions” or “incubations” are performed *in vitro*, high-stringency conditions typically comprise lower salt concentrations, a temperature of 65C or greater, and a detergent, such as sodium dodecylsulfate (SDS) at a concentration ranging from about 0.1% to about 2%. When DNA-protein “binding reactions” or “incubations” are performed within living cells, the stringency of the binding reaction is controlled as described by Joung et al. (Joung et al., 2000, Proceedings of the National Academy of Sciences (USA) 97:7382 and US Patent Application No. 20020119498).

Further definitions are provided in context below.

III. NRSF as a Scaffold for Production of Synthetic Multi-Finger Zinc Finger Proteins

The methods of the present invention involve engineering the DNA binding specificity of the naturally occurring Zf protein, NRSF, to produce non-naturally occurring NRSF-derived Zf proteins capable of binding to any extended DNA target sequence of interest.

The full length human NRSF protein consists of 1097 amino acid residues. The amino acid sequence of this protein, which was first reported by Chong et al. (Cell 80 (6) 949-957 (1995)) and Schoenherr et al. (Science 267 (5202) 1360-1363 (1995)), is provided in Figure 1 (,) and has Gene Bank accession no. NP_005603. The cDNA encoding the human NRSF protein consists of 3294 bases. The nucleotide sequence of this cDNA is provided in Figure 2 (SEQ ID NO. 2) and has Gene Bank accession no. NM_005612. The amino acid and nucleotide sequences of mouse (Figure 3 & 4, SEQ ID NO. 3 & 4), rat (Figure 5 & 6, SEQ ID NO. 5 & 6), and Xenopus (Figure 7 & 8, SEQ ID NO. 7 & 8) are also illustrated. Figure 9 illustrates the amino acid sequences of each of the 8 zinc fingers and the inter-finger linkers of human NRSF (taken from SEQ ID NO. 1). U.S. patents 5,935,811 and 6,270,990 describe the nucleotide and amino acid sequence of the wild-type human NRSF protein.

As with all C₂H₂-type Zf proteins, the Zfs in the NRSF DNA binding domain contain 2 conserved cysteine residues and 2 conserved histidine residues. However, in several respects, the DNA binding domain of NRSF differs from those found in most other Zf transcription factors. For example, the Zfs in NRSF most closely match the less common zinc finger consensus sequence Y-X-C-X₂-C-X-F-X₇-L-X₂-H-X₄-H (SEQ ID NO. 11), as opposed to the more common motif (F/Y)-X-C-X_{2.5}-C-X₃-(F/Y)-X₅-Φ-X₂-H-X_{3.5}-H (SEQ ID NO. 12) (where X is any amino acid and Φ is a hydrophobic amino acid) found in proteins such as Zif268. All 8 Zfs of NRSF harbor a tyrosine residue at the position that is two amino acids carboxy-terminal to the second conserved cysteine (see Figure 2 above) as opposed to the more commonly found phenylalanine. In addition, several of the inter-finger linkers in NRSF differ in length and/or composition from the consensus TGEKP-type linkers found in many other Zf proteins including zif268. For example, the linkers between Zf1 and Zf2 and between Zf2 and Zf3 comprise 34 and 9 amino acids, respectively. Also, although Zfs 3 to 8 are all connected by 5 amino acid inter-finger linkers, only 2 of these (the Zf4-Zf5 and Zf5-Zf6 linkers) are of the TGEKP type. Without being bound by theory, it is believed that some or all of these unusual characteristics of

the NRSF DNA binding domain may provide NRSF with its unique capability to bind simultaneously to each of 20 base pairs within the NRSE target sequence.

To alter the DNA binding specificity of the human NRSF protein according to the methods of the present invention, the amino acid sequence of the zinc fingers in the DNA binding domain are altered. Any suitable method known in the art can be used to alter the amino acid sequence of the NRSF Zfs, such as random mutagenesis, PCR, synthetic construction and the like. (see, e.g., U.S. Pat. No. 5,786,538; Wu et al., PNAS 92:344-348 (1995); Jamieson et al., Biochemistry 33:5689-5695 (1994); Rebar & Pabo, Science 263:671-673 (1994); Choo & Klug, PNAS 91:11163-11167 (1994); Choo & Klug, PNAS 91: 11168-11172 (1994); Desjarlais & Berg, PNAS 90:2256-2260 (1993); Desjarlais & Berg, PNAS 89:7345-7349 (1992); Pomerantz et al., Science 267:93-96 (1995); Pomerantz et al., PNAS 92:9752-9756 (1995); and Liu et al., PNAS 94:5525-5530 (1997); Griesman & Pabo, Science 275:657-661 (1997); Desjarlais & Berg, PNAS 91:11-99-11103 (1994), Joung et al., PNAS (2000)). In a preferred aspect, the amino acid sequences of the zinc fingers are altered randomly to generate combinatorial libraries of sequences derived from NRSF. Methods for randomization of amino acid sequences and for production of libraries encoding such randomized peptides are routine practice to those skilled in the art, and any such method can be used to produce randomized NRSF-based libraries. Preferred libraries and selection strategies are described below.

In the most general embodiment of the present invention, the amino acid sequence of the NRSF DNA binding domain can be altered in any way desired, the only requirement being that the DNA binding specificity is altered such that the new NRSF-derived protein binds to the target sequence of interest while the wild-type NRSF DNA binding domain does not bind to the target sequence of interest. This may be achieved by altering anywhere from one amino acid in one zinc finger to all of the amino acids within each of the 8 zinc fingers. Also, this might be achieved by altering the amino acid sequence of the linkers that connect each of the zinc fingers in the NRSF DNA binding domain.

However, in a preferred embodiment the amino acid alterations that are made to the NRSF DNA binding domain are constrained such that certain of the unique features of the NRSF DNA binding domain described above are retained. For example, it is preferred that the engineered protein has a tyrosine residue at the position that is two amino acids carboxy-terminal to the second conserved cysteine in each zinc finger. It is also preferred that the NRSF-derived-

protein retains approximately the same number of amino acid residues in the inter-finger linkers as occur in the WT NRSF protein (see Figure 9). Thus, in a preferred embodiment the inter-finger linker between Zf 1 and Zf 2 comprises about 34 amino acid residues, the linker between Zf 2 and Zf3 comprises about 9 amino acid residues, and the remaining inter-finger linkers (i.e. those between Zf3 and Zf4, Zf4 and Zf5, Zf5 and Zf6, Zf6 and Zf7, Zf7 and Zf8) are approximately 5 amino acids in length. In an even more preferable embodiment, the linkers between each Zf in the engineered NRSF-derived protein have the same amino acid sequence as the inter-finger linkers in the WT NRSF protein.

IV. NRSF-based libraries and strategies for selection of NRSF-based Zf proteins

Any strategy suitable for selection of multi-finger proteins can be used for the selection of a NRSF-based Zf protein. For a review of some of such methods see Beerli and Barbas, (2002) *Nature Biotechnology* 20:135. For example, suitable selection strategies include Greisman and Pabo's "sequential selection" method (Greisman and Pabo (1997) *Science* 275:657 and US Patent No. 6,410,248), the "bipartite selection" method developed by Isalan et al. (Isalan et al., (2001) *Nature Biotechnology* 19: 656), and the "parallel selection" methods described by Desjarlais et al., (Proceedings of the National Academy of Sciences (USA) 90:2256, (1993)) and Choo et al., (*Nature* 372:642, (1994)).

However, in a preferred embodiment the "Context Sensitive Parallel Optimization" or "CSPO" strategy developed by Joung et al., is used to select the NRSF-based Zf proteins of the present invention. The general principles and detailed methods of the CSPO strategy are described in U.S. Provisional Patent Application Serial No. 60/420,458, the contents of which are hereby incorporated by reference. The specific application of CSPO to the selection of NRSF-based Zf proteins is described below.

Any suitable expression system can be used for expression of NRSF-based libraries for example, phage display (see U.S. Patent No. 6,013,453 and U.S. Patent No. 6,007,988), polysome display (WO 0027878 A1), *in vitro* transcription/translation, or expression in eukaryotic or prokaryotic cells, methods for which are well known in the art. Likewise, any suitable selection methods can be used to select those expressed NRSF-based Zf proteins in the library that have the desired DNA binding characteristics. In a preferred embodiment, a eukaryotic or prokaryotic cell-based system is used for both expression of the NRSF-based libraries and the selection of the NRSF-based proteins that bind to the target sequence of interest. The use of such a cell-based system

advantageously provides for the selection Zf proteins that are likely to function well in a cellular context. In the most preferred embodiment, a bacterial "2-hybrid" system is used to express and select the Zfs of the present invention. The bacterial 2-hybrid selection method has an additional advantage, in that the library protein expression and the DNA binding "assay" occur within the same cells. The use of bacterial 2-hybrid systems to express and select Zf proteins is described in Joung et al., 2000, Proceedings of the National Academy of Sciences (USA) 97:7382 and US Patent Application No. 20020119498, the contents of which are incorporated herein by reference.

V. Selection of the Target Sequence of Interest

As described herein, Zfs can be designed to recognize any suitable target sequence. Thus, any target sequence in any gene of interest can be chosen, and used as the "template" against which to select a NRSF-based Zf protein. A general theme in transcription factor function is that simple binding and sufficient proximity to the promoter are all that is generally needed. Therefore, the exact positioning of the chosen target site relative to the promoter (both in terms of orientation and, within limits, distance) do not matter greatly. This allows considerable flexibility in choosing target sites. The target site recognized by the NRSF-based Zf can be any suitable site in the target gene that will allow regulation of gene expression by a NRSF-based Zf, optionally linked to a regulatory domain. Preferred target sites include regions adjacent to, downstream, or upstream of the transcription start site. In addition, target sites that are located in enhancer regions, repressor sites, RNA polymerase pause sites, and specific regulatory sites (e.g., SP-1 sites, hypoxia response elements, nuclear receptor recognition elements, p53 binding sites), sites in the cDNA encoding region or in an expressed sequence tag (EST) coding region, can be used.

In a preferred embodiment, the target site is chosen such that the sequence is statistically unique enough to occur only once in the genome. This ability to specify a unique sequence is a function of the length of the target site and the size of the genome or other desired substrate (such as a nucleic acid vector, for example). For example, assuming random base distribution, a unique 16 bp sequence will occur only once in 4.3×10^9 bp, thus a 16 bp sequence should be sufficient to specify a unique address within 4.3×10^9 bp of sequence. Similarly, an 18 bp address would enable sequence specific targeting within 6.8×10^{10} bp of DNA. However, it should be noted that the "effective" frequency of such unique addresses in the human genome is

likely to be significantly lower than the frequencies predicted by these purely statistical calculations, because a certain portion of the DNA in the genome is packaged into regions of densely packed chromatin that is not accessible by transcription factors. The unique target site selected can be located anywhere within or proximal to the gene of interest. Wherein the ultimate aim is to generate a synthetic transcription factor to regulate expression of the gene of interest, it is preferable that the chosen target site is within the general vicinity of the promoter and in a region where chromatin architecture will not impede binding of the Zf protein to the target site (see for example, Liu et al., (2001) Journal of Biological Chemistry 276:11323).

Although any potential target site sequence can be used, the present invention provides methods for predicting and selecting target sequences that are most likely to provide good substrates against which to select an NRSF-based Zf protein. The methods provided by the present invention for selection of optimum target site make use of knowledge gained about the details of the NRSF-NRSE interaction. Modeling of the specific base contacts made by each of the Zfs within NRSF (see Example 3) has enabled the development of a "framework" sequence, a partially degenerate version of the 21 base pair consensus NRSE (e.g. ^{5'}NNNNN(C/G)NNCNCNNGNNCNN^{3'} SEQ ID NO. 13) that limits the possible sequences chosen as a target site. It is preferred that all potential target sequences chosen match this framework sequence. The fixed, non-degenerate bases in any framework sequence will be those that are contacted by recognition helix residues from more than one finger at the NRSF-NRSE interface. This limitation stems from the fact that alteration of one of these "finger overlap" bases might require randomization of more than one finger to recognize a new base at that position..

Target sites can be chosen in any gene or other nucleotide sequence (such as vectors, plasmids etc.) desired. Examples of endogenous genes suitable for regulation include VEGF, erbB2, CCR5, ER.alpha., Her2/Neu, Tat, Rev, HBV C, S, X, and P, LDL-R, PEPCK, CYP7, Fibrinogen, ApoB, Apo E, Apo(a), renin, NF-.kappa.B, I-.kappa.B, TNF-.alpha., FAS ligand, amyloid precursor protein, atrial natriuretic factor, ob-leptin, ucp-1, IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-12, G-CSF, GM-CSF, Epo, PDGF, PAF, p53, Rb, fetal hemoglobin, dystrophin, eutrophin, GDNF, NGF, IGF-1, VEGF receptors fit and flk, topoisomerase, telomerase, bcl-2, cyclins, angiostatin, IGF, ICAM-1, STATS, c-myc, c-myb, TH, PTI-1, polygalacturonase, EPSP synthase, FAD2-1, delta-12 desaturase, delta-9 desaturase, delta-15 desaturase, acetyl-CoA

carboxylase, acyl-ACP-thioesterase, ADP-glucose pyrophosphorylase, starch synthase, cellulose synthase, sucrose synthase, senescence-associated genes, heavy metal chelators, fatty acid hydroperoxide lyase, viral genes, protozoal genes, fungal genes, and bacterial genes. In general, suitable genes to be regulated include cytokines, lymphokines, growth factors, mitogenic factors, chemotactic factors, onco-active factors, receptors, potassium channels, G-proteins, signal transduction molecules, and other disease-related genes.

VI. CSPO Method for Selection of NRSF-Based ZF Proteins

i. NRSF-based Primary Libraries

CSPO is an efficient Zf selection strategy that allows assembled multi-finger polypeptides to be selected for binding to a desired sequence of interest while also retaining maximal combinatorial diversity in the Zf libraries used. Zf polypeptides identified using CSPO typically have an affinity and specificity for their target site that is superior to that produced by alternative methods. The CSPO method involves the 2 sequentially performed selection steps using 2 sets of libraries.

A separate primary library must be used for each Zf position within the multi-finger protein to be generated. Thus, to select an 8 finger NRSF-based Zf protein, 8 different primary libraries are produced. The first primary library has Zf position 1 (the N-terminal Zf) randomized and Zf positions 2-8 held constant as "anchor" fingers. The second primary library has Zf position 2 (the finger C-terminal to Zf position 1) randomized and Zf positions 1 and 3-8 held constant as "anchor" fingers. The third primary library has Zf position 3 (the finger C-terminal to Zf position 2) randomized and Zf positions 1, 2 and 4-8 held constant as "anchor" fingers, and so on.

Similarly, the same general method can be used to select NRSF-based Zf proteins having fewer than 8 Zfs. For example, a 4 Zf protein can be derived from an NRSF scaffold by using just 4 primary libraries, each having one different NRSF finger randomized and having 3 constant anchor fingers. Similarly, a 6 Zf protein can be derived from an NRSF scaffold by using 6 primary libraries, each having one different NRSF finger randomized and having 5 constant anchor fingers.

Importantly, it has been surprisingly shown that Zfs 3-8 of NRSF alone are sufficient to bind to a 20 bp NRSE target sequence, and that Zfs 1 and 2 of NRSF are not required for this binding. Therefore, in a preferred embodiment generation of a NRSF-based Zf protein for

binding to a desired target sequence of 20 bp or less is performed using a maximum of 6 primary libraries, each having one of Zfs 3-8 varied. Although fingers 1 and 2 of NRSF are not required for binding to the consensus NRSE, they may make indirect contributions to DNA binding affinity and specificity (see Example 4). Therefore, it is preferred that Zfs 1 and 2 are retained in these NRSF-based libraries but are simply not varied. This will result in the selection of eight-finger proteins, but only 6 of the Zfs within the eight-finger protein will have been "engineered". However, if desired either or both of Zf 1 and Zf 2 can be deleted from the NRSF-based libraries.

The constant "anchor" fingers in the primary libraries can be any zinc fingers chosen from any known zinc finger protein. In one embodiment the "anchor" fingers are those of the wild-type NRSF protein.

In a preferred embodiment 6 amino acid residues within a single zinc finger are randomized. In a still more preferred embodiment the 6 amino acid residues randomized within a given zinc finger are the amino acid residues at positions -1, 1, 2, 3, 5, and 6, where position 1 is the first residue of the α -helical section of each zinc finger (see Figure 9).

The number of randomized amino acids at a single variable residue position can be varied up to the maximum limits of the library expression and selection system used. Preferably, all 20 naturally occurring amino acids are represented at all randomized residue positions. However, more frequently, it will be desirable to limit the number of amino acids represented at any given residue position to 19. If cysteine is excluded, the remaining 19 naturally occurring amino acids can be encoded by 24 codons as a result of codon doping schemes (Wolfe et al., (2001) *Structure* 9:717). Libraries with 24 codon variations at 6 variable positions of an α -helix have a diversity of 24^6 . A library of such a size is within the limits of known expression and selection systems, such as the bacterial 2-hybrid system and phage display. Thus, in one embodiment, methods of the present invention comprise the use of libraries in which 19 different naturally occurring amino acids are represented at one or more variable residue positions of the α -helix. In this instance, the naturally occurring amino acid cysteine is excluded because cysteine can not readily be incorporated into a 24-codon doping strategy.

In yet another embodiment, 16 naturally occurring amino acids are represented in any given randomized residue position within the α -helix. 16 amino acids can also be encoded by 24 codons using codon-doping strategies (see Joung et al., (2000) *Proceedings of the National*

Academy of Sciences (USA) 97:7382). Thus, as for the 19 amino acid library described above, such a 16 amino acid Zf library also has a diversity of 24^6 . In the embodiment where a 16 amino acid/24 codon library is used, the excluded amino acids are preferably phenylalanine, tryptophan, tyrosine, and cysteine.

The primary libraries described herein can be synthesized using any known randomization strategy (see for example Joung et al., (2000) Proceedings of the National Academy of Sciences (USA) 97:7382), U.S. Patent No. 6,013,453 and U.S. Patent No. 6,007,988). Such strategies are well known to those skilled in the art and include, for example, the use of degenerate oligonucleotides, use of mutagenic cassettes and techniques based on error prone PCR. Standard recombinant DNA and cloning techniques can also be used for library construction and for incorporation of such libraries into appropriate expression and selection systems. Standard recombinant DNA and cloning techniques are well known to those of skill in the art and are described in laboratory text such as, for example, Sambrook et al., Molecular Cloning; A Laboratory Manual 2d ed. (1989), the contents of which are incorporated herein by reference.

ii. Target site constructs for use in primary screen.

Once the desired target site in the gene or sequence of interest has been chosen, "target site constructs" for use in screening assays can be produced. The CSPO strategy employs construction and/or use of a separate "target site construct" for each subsite within the entire target site. For example, if an 18 bp (6 subsite) target site is chosen, 6 "target site constructs" are produced. In the first "target site construct" subsite 1 (the 5' subsite) would have the sequence of the gene of interest, and subsites 2-6 would have defined "anchor" sequences. These anchor sequences are the sequence bound by the "anchor fingers" described above. In the second target site construct subsite 2 would have the sequence of the gene of interest, and subsites 1 and 3-6 would have the defined "anchor" sequences. In summary only one subsite within a given "target site construct" should have the sequence of the gene of interest and the remaining subsites should have the defined "anchor" sequences. These target sites are referred to as "position sensitive" because the subsites having the sequence of the gene of interest are located at the same position relative to the other subsites, as occurs in the true target site within the gene of interest.

In a preferred embodiment, these "target site constructs" are cloned upstream of a test promoter in a vector for use in the bacterial 2-hybrid system (Joung et al., 2000, Proceedings of

the National Academy of Sciences (USA) 97:7382 and US Patent Application No. 2002011949. Such target site constructs can be synthesized readily using standard molecular biology techniques (for example using restriction digestion of vector DNA, PCR, or automated nucleic acid synthesis). Such techniques are well known to those skilled in the art and are described in many laboratory texts such as, for example Sambrook et al., *Molecular Cloning, A Laboratory Manual* 2d ed. (1989).

iii. Primary Selection

A key feature of the CSPO Zf selection strategy is that a separate primary selection is performed for each "Zf/subsite pair" i.e. if the aim is to select a 6 finger NRSF-based protein that binds to a 20 bp target sequence, 6 parallel primary selections are performed, one for each randomized finger. For example, in the scheme described above, in primary selection 1, primary library 1 is expressed and candidates are selected for binding to DNA target site 1, i.e. primary library 1 and DNA target site 1 comprise a Zf/subsite pair. Similarly, in primary selection 2, primary library 2 is expressed and candidates are selected for binding to DNA target site 2.

In a preferred embodiment, the stringency of each of the primary selections should be low, such that each selection yields a pool of NRSF-based Zf proteins with target binding affinities that range from low to high. The rationale for this low stringency selection is that there should be no bias towards Zfs that bind tightly to their target subsite at the primary selection stage, because Zfs so identified may not bind tightly to their target subsite in the context of the Zfs selected against the other subsites that make up the full target sequence. Zfs that bind tightly in the context of the "anchor" fingers may not bind tightly in the context of the full target specific Zf protein. Mechanisms for controlling the stringency of DNA binding reactions are known to those of skill in the art and any such mechanism can be used.

iv. Construction of Secondary Partially Optimized Library

The primary screening methods described above will yield a separate "pool" of candidate NRSF-based Zf proteins for each "Zf/subsite" pair. A key aspect of the CSPO strategy is that these "pools" can be recombined to produce a secondary library comprising variants that harbor fingers which have been partially optimized for binding to a desired subsite. For example, such a secondary library can comprise a range of multi-finger proteins composed of random combinations of the pools of fingers selected from the randomized fingers of the primary library. Thus, the secondary library can comprise multi-finger proteins that, unlike the primary library,

can potentially vary at all finger positions of the multi-finger proteins. Furthermore, the secondary library can comprise fingers with a range of binding affinities and specificities for their target subsite(s). The secondary library can then be used in a secondary screen, which is preferably conducted under conditions of high-stringency, to produce a multi-Zf polypeptide that binds with high affinity to the sequence of interest. Preferably, a new secondary library is synthesized for each new multi-finger protein to be produced.

The individual "pools" derived from the individual primary selections can be recombined using any one of a number of recombination techniques known in the art, such as described in, for example, Sambrook et al., *Molecular Cloning; A Laboratory Manual* 2d ed. (1989). Preferably, the individual "pools" derived from the individual primary selections are recombined using a PCR-mediated recombination method. More preferably still, the individual "pools" derived from the individual primary selections are recombined using a PCR-mediated recombination method.

v. Secondary Selection

For each new sequence specific NRSF-based Zf protein to be produced, a single high-stringency secondary screen is performed. In this screen, a partially optimized secondary library (such as described above) is screened against the exact target sequence of interest, wherein the sequence of interest excludes "anchor" subsites. Thus, in the secondary screen, full-length assembled NRSF-based Zfs that bind to the sequence of interest can be identified. This is a key feature of the CSPO strategy, and means that there is no need to perform any post-selection assembly of individual Zfs or groups of Zfs to generate the final multi-finger product. Such post-selection assembly is a common feature of other Zf selection methods. Post-selection assembly of 10 introduces an uncontrollable element into the production of multi-finger proteins, as there is a possibility that the individually selected fingers will not function as predicted when assembled into the final multi-finger protein. CSPO advantageously allows for secondary selection of fully assembled NRSF-based Zfs and thus results in the generation of a final product where each NRSF-based finger and the linkers between them (also NRSF-based) are known to work together to bind to the target sequence of interest. In a preferred embodiment, the secondary selection is performed at high-stringency in order to isolate proteins that bind to their sequence of interest with high affinity. Mechanisms for controlling the stringency of selection reactions are known to those of skill in the art and any such mechanism can be used.

VII. Characterization of Selected NRSF-based Proteins.

The NRSF-based Zf proteins identified using methods of the present invention can be further characterized after selection to ensure that they bind to the target site of interest with the desired characteristics, and to confirm that the selected NRSF-based proteins do not bind to other related or similar sequences. It is preferred that any selected NRSF-based proteins that do not bind to the target sequence with high specificity should be eliminated from subsequent development. It is preferred that the selected proteins be tested for target site binding using a different strategy than that used in the original selection, thereby controlling for the possibility of spurious or artifactual interactions specific to the selection system. For example, Zfs selected using a bacterial 2-hybrid or phage-display system can be assayed for binding to their target sequence using an electrophoretic mobility shift assay or "EMSA" (Buratowski & Chodosh, in *Current Protocols in Molecular Biology* pp. 12.2.1-12.2.7). Equally, any other DNA binding assay known in the art could be used to verify the DNA binding properties of the selected NTSF-based protein.

Preferably, calculations of binding affinity and specificity are also made. This can be done by a variety of methods. The affinity with which the selected NRSF-based Zf protein binds to the sequence of interest can be measured and quantified in terms of its K_D . Any assay system can be used, as long as it gives an accurate measurement of the actual K_D of the Zf protein. In one embodiment, the K_D for the binding of a Zf protein to its target is measured using an EMSA.

In a preferred embodiment, EMSA is used to determine the K_D for binding of the selected Zf protein both to the sequence of interest (i.e. the specific K_D) and to non-specific DNA (i.e. the non-specific K_D). Any suitable non-specific or "competitor" double stranded DNA known in the art can be used. Preferably, calf thymus DNA or human placental is used. The ratio of the specific K_D to the non-specific K_D can be calculated to give the specificity ratio. Zfs that bind with high specificity have a high specificity ratio. This measurement is very useful in deciding which of a group of selected Zfs should be used for a given purpose. For example, use of Zfs *in vivo* requires not only high affinity binding but also highly-specificity binding. In a preferred embodiment, Zfs isolated using methods of the present invention have binding specificities higher than Zfs selected using other selection strategies (such a parallel selection, sequential selection and bipartite selection), and even more preferably, comparable or superior to those of naturally occurring multi-finger proteins, such as Zif268. It is preferred that the NRSF-based

proteins of the present invention bind to their target sequences with affinities in the picomolar to nanomolar range. Furthermore, it is preferred that the NRSF-based proteins of the present invention bind to their target with a specificity ratio of $>250,000$ (i.e. the ratio one would expect for a perfectly specific three-finger protein that specifies 9 base pairs of DNA). Further methods useful for the characterization of selected NRSF-based Zf proteins *in vitro* and in living cells are provided in Examples 8 and 9, respectively.

VIII. Construction of NRSF-based Transcription Factors.

The ultimate aim of producing a NRSF-based Zf protein is to obtain a Zf that is able to specifically bind to an extend DNA sequence, and that can be used to perform a function. The NRSF-based proteins of the present invention can be used to perform any of the functions already described for other types of synthetic zinc finger molecules. For example, the NRSF-based Zf protein can be used alone, for example to bind to a specific site on a gene and thus block binding of other transcription factors. In a preferred embodiment, the NRSF-based Zf is used in the construction of a recombinant NRSF-based transcription factor which can be used for a variety of purposes including regulation gene expression *in vivo* for the treatment of disease (see, for example, U.S. patent application 2002/0160940 A1, and U.S. Patent Nos. 6,511,808, 6,013,453 and 6,007,988, and International patent application WO 02057308 A2), or for performing either *in vivo* or *in vitro* functional genomics studies (see, for example, U.S. Patent No. 6,503,717 and U.S. patent application 2002/0164575 A1). To generate a functional transcription factor, an NRSF-based Zf "DNA binding domain" is fused to an "effector" domain. Fusing effector domains to synthetic Zf proteins to form functional transcription factors involves only routine molecular biology techniques which are routinely practiced by those of skill in the art, see for example, U.S. Patent Nos. 6,511,808, 6,013,453, 6,007,988, 6,503,717 and U.S. patent application 2002/0160940 A1).

In one embodiment the DNA binding domain used to form the synthetic transcription factor of the present invention is the exact NRSF-based protein that has been selected. In other embodiments, two or more selected NRSF-based Zf proteins are linked together to produce the final DNA binding domain. The linkage of two or more selected NRSF-based proteins may be performed by covalent or non-covalent means. In the case of covalent linkage NRSF-based proteins can be covalently linked together using an amino acid linker (see, for example, U.S. patent application 2002/0160940 A1, and International applications WO 02099084A2 and WO

0153480 A1). This linker may be any string of amino acids desired. In one embodiment the linker is a canonical TGEKP linker. In a preferred embodiment the linker has the same sequence as one of the linkers in the wild-type NRSF protein. Whatever linkers are used standard recombinant DNA techniques (such as described in, for example, Sambrook et al., *Molecular Cloning; A Laboratory Manual* 2d ed. (1989)) are used to produce such linked proteins.

In the case of non-covalent linkage, two or more NRSF-based proteins are multimerized. Where only two NRSF-based proteins are non-covalently linked the proteins are said to be dimerized. In one embodiment two identical NRSF-based proteins may be linked to form a homo-dimer. In an alternative embodiment two different NRSF-based proteins may be linked to form a hetero-dimer. For example, a six-finger protein may be produced by dimerization of two three-finger proteins, or an eight-finger protein may be produced by dimerization of two four-finger proteins. The production of multimers or dimers can be performed by fusing "multimerization" or "dimerization domains" to the zinc finger proteins to be joined. Any suitable method for fusing protein domains or producing chimeric proteins can be used. For example, in one embodiment, the DNA encoding the zinc finger protein is fused to the DNA encoding the multimerization domain using standard recombinant DNA techniques (as described in, for Example, Sambrook et al., *Molecular Cloning; A Laboratory Manual* 2d ed. (1989)).

Suitable multimerization or dimerization domains can be selected from any protein that is known to exist as a multimer or dimer, or any protein known to possess such multimerization or dimerization activity. Examples of suitable domains include the dimerization element of Gal4, leucine zipper domains, STAT protein N-terminal domains, and FK506 binding proteins (see, e.g., Pomerantz et al., *Biochemistry* 37: 965-970 (1998), Wolfe et al., *Structure* 8: 739-750 (2000), O'Shea, *Science* 254: 539 (1991), Barahmand-Pour et al., *Curr. Top. Microbiol. Immunol.* 211:121-128 (1996); Klemm et al., *Annu. Rev. Immunol.* 16:569-592 (1998); Ho et al., *Nature* 382:822-826 (1996)). Furthermore, some zinc finger proteins themselves have dimerization activity. For example, the zinc fingers from the transcription factor Ikaros have dimerization activity (McCarty et al., *Molecular Cell* 11: 459-470 (2003)), and there is evidence to suggest that even the zinc fingers of NRSF (and/or NRSF splice variants) may have some dimerization activity (Shimojo et al., *Mol Cell Biol.* 19: 6788-95 (1999)). Thus, in one embodiment zinc fingers having dimerization function are used. In the event that the zinc fingers in the selected NRSF-based proteins themselves have dimerization function then there will be no

need to fuse an additional dimerization domain to these proteins. In certain embodiments, "conditional" multimerization of dimerization" technology can be used. For example, this can be accomplished using FK506 and FKBP interactions. FK506 binding domains are attached to the proteins to be dimerized. These proteins will remain apart in the absence of a dimerizer. Upon addition of a dimerizer, such as the synthetic ligand FK1012, the two proteins will fuse.

A particular advantage of the dimerization methods of the present invention, is that because of the low DNA-binding affinity of the NRSF zinc fingers (e.g. three fingers of NRSF bind with an affinity in the micromolar range whereas three fingers of Zif268 bind with an affinity in the picomolar to nanomolar range) the number of zinc fingers that can be linked together is not limited. For example, using the methods of the present invention two three-finger NRSF-based proteins may be dimerized to produce a six-finger protein, two four-finger proteins may be dimerized to produce an eight finger protein, or two five-finger proteins may be linked together to produce a ten-finger protein. This is in contrast to dimerization of zif286-based fingers where it has only been possible to link two two-finger proteins together to form a four-finger protein (Wolfe et al., Structure 8: 739-750 (2000)). In the case of zif268 dimerization of larger zinc finger domains results in the production of a protein whose DNA-binding affinity is likely to be too high for the protein to be physiologically useful (Wolfe et al., Structure 8: 739-750 (2000)).

The "effector" domain can be associated with the Zf protein at any suitable position, including the C- or N-terminus of the Zf protein. Suitable "effector" domains for addition to the NRSF-based protein made using the methods of the invention are described in U.S. Patent Nos. 6,511,808, 6,013,453, 6,007,988, U.S. and 6,503,717 and U.S. patent application 2002/0160940 A1. Such effector domains include transcription factors (activators, repressors, co-activators, co-repressors), silencers, nuclear hormone receptors, oncogene transcription factors (e.g., myc, jun, fos, myb, max, mad, rel, ets, bcl, myb, mos family members etc.), restriction endonucleases (e.g.—the FokI restriction enzyme; see Kim et al. PNAS (1996) Feb 6;93(3):1156-60); and chromatin associated proteins and their modifiers (e.g. methylases, demethylases, acetylases and deacetylases). Kinases, phosphatases, and other proteins that modify polypeptides involved in gene regulation are also useful as regulatory domains for Zf proteins. Such modifiers are often involved in switching on or off transcription mediated by, for example, hormones. Kinases

involved in transcription regulation are reviewed in Davis, Mol. Reprod. Dev. 42:459-67 (1995). Phosphatases are reviewed in, for example, Schonthal & Semin, Cancer Biol. 6:239-48 (1995).

In a preferred embodiment the regulatory domains found in the native NRSF protein are used to generate the final synthetic NRSF-based transcription factor. The native NRSF protein comprises an N-terminal repressor domain and a C-terminal repressor domain (Tapia Ramirez et al., 1997 PNAS 94; p1172-1182; Andres et al., 1999 PNAS 96; p9873-9878; Grimes et al., 2000 Journal of Biological Chemistry 275: p9461-9467). Either or both of these repressor domains may be used. It has recently been shown that the C-terminal repressor domain of NRSF can mediate long term silencing through alteration of chromatin structure (Lunyak et al., 2002 Science 298; p1747-1751). Thus, it may be particularly desirable to use NRSF-based Zf proteins comprising the C-terminal repressor domain of NRSF in circumstances where long-term or permanent "switching-off" of the target gene is desired. Another advantage of using the C-terminal repressor domain of NRSF is that target cells may only need to be exposed to the NRSF-based protein briefly to result in long term regulation of gene expression. This will be particularly useful in human patients, as it means a single short term "treatment" with such a NRSF-based protein may be all that is required to induce long term effects on gene expression.

Fusions of NRSF-selected Zfs to regulatory domains can be performed by standard recombinant DNA techniques well known to those skilled in the art, and as are described in, for example, basic laboratory texts such as Sambrook et al., Molecular Cloning; A Laboratory Manual 2d ed. (1989), and in U.S. Patent Nos. 6,511,808, 6,013,453, 6,007,988, U.S. and 6,503,717 and U.S. patent application 2002/0160940 A1.

IX. Use of Selected NRSF-based Proteins

The ultimate aim of producing NRSF-based transcription factors is to express and produce the NRSF-based proteins, and use them to regulate gene expression, either *in vitro* or *in vivo*. Further description of how this can be achieved are given below

i. Expression Vectors

The nucleic acid encoding the NRSF-based Zf protein is typically cloned into intermediate vectors for transformation into prokaryotic or eukaryotic cells for replication and/or expression. Intermediate vectors are typically prokaryote vectors, e.g., plasmids, or shuttle vectors, or insect vectors, for storage or manipulation of the nucleic acid encoding the NRSF-based Zf protein or production of protein. The nucleic acid encoding the NRSF-based Zf protein

is also typically cloned into an expression vector, for administration to a plant cell, animal cell, preferably a mammalian cell or a human cell, fungal cell, bacterial cell, or protozoal cell.

To obtain expression of a cloned gene or nucleic acid, the NRSF-based Zf protein is typically subcloned into an expression vector that contains a promoter to direct transcription. Suitable bacterial and eukaryotic promoters are well known in the art and described, e.g., in Sambrook et al., *Molecular Cloning, A Laboratory Manual* (2nd ed. 1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology* (Ausubel et al., eds., 1994). Bacterial expression systems for expressing the NRSF-based Zf protein are available in, e.g., *E. coli*, *Bacillus* sp., and *Salmonella* (Palva et al., *Gene* 22:229-235 (1983)). Kits for such expression systems are commercially available. Eukaryotic expression systems for mammalian cells, yeast, and insect cells are well known in the art and are also commercially available.

The promoter used to direct expression of the NRSF-based Zf protein nucleic acid depends on the particular application. For example, a strong constitutive promoter is typically used for expression and purification of the NRSF-based Zf protein. In contrast, when the NRSF-based Zf protein is to be administered *in vivo* for gene regulation, either a constitutive or an inducible promoter is used, depending on the particular use of the NRSF-based Zf protein. In addition, a preferred promoter for administration of the NRSF-based Zf protein can be a weak promoter, such as HSV TK or a promoter having similar activity. The promoter typically can also include elements that are responsive to transactivation, e.g., hypoxia response elements, Gal4 response elements, lac repressor response element, and small molecule control systems such as tet-regulated systems and the RU-486 system (see, e.g., Gossen & Bujard, *PNAS* 89:5547 (1992); Oligino et al., *Gene Ther.* 5:491-496 (1998); Wang et al., *Gene Ther.* 4:432-441 (1997); Neering et al., *Blood* 88:1147-1155 (1996); and Rendahl et al., *Nat. Biotechnol.* 16:757-761 (1998)).

In addition to the promoter, the expression vector typically contains a transcription unit or expression cassette that contains all the additional elements required for the expression of the nucleic acid in host cells, either prokaryotic or eukaryotic. A typical expression cassette thus contains a promoter operably linked, e.g., to the nucleic acid sequence encoding the ZFP, and signals required, e.g., for efficient polyadenylation of the transcript, transcriptional termination,

ribosome binding sites, or translation termination. Additional elements of the cassette may include, e.g., enhancers, and heterologous spliced intronic signals.

The particular expression vector used to transport the genetic information into the cell is selected with regard to the intended use of the NRSF-based Zf protein, e.g., expression in plants, animals, bacteria, fungus, protozoa etc. (see expression vectors described below and in the Example section). Standard bacterial expression vectors include plasmids such as pBR322 based plasmids, pSKF, pET23D, and commercially available fusion expression systems such as GST and LacZ. A preferred fusion protein is the maltose binding protein, "MBP." Such fusion proteins are used for purification of the NRSF-based Zf protein. Epitope tags can also be added to recombinant proteins to provide convenient methods of isolation, for monitoring expression, and for monitoring cellular and subcellular localization, e.g., c-myc or FLAG.

Expression vectors containing regulatory elements from eukaryotic viruses are often used in eukaryotic expression vectors, e.g., SV40 vectors, papilloma virus vectors, and vectors derived from Epstein-Barr virus. Other exemplary eukaryotic vectors include pMSG, pAV009/A+, pMTO10/A+, pMAMneo-5, baculovirus pDSVE, and any other vector allowing expression of proteins under the direction of the SV40 early promoter, SV40 late promoter, metallothionein promoter, murine mammary tumor virus promoter, Rous sarcoma virus promoter, polyhedrin promoter, or other promoters shown effective for expression in eukaryotic cells.

Some expression systems have markers for selection of stably transfected cell lines such as thymidine kinase, hygromycin B phosphotransferase, and dihydrofolate reductase. High yield expression systems are also suitable, such as using a baculovirus vector in insect cells, with the NRSF-based Zf protein encoding sequence under the direction of the polyhedrin promoter or other strong baculovirus promoters.

The elements that are typically included in expression vectors also include a replicon that functions in *E. coli*, a gene encoding antibiotic resistance to permit selection of bacteria that harbor recombinant plasmids, and unique restriction sites in nonessential regions of the plasmid to allow insertion of recombinant sequences.

Standard transfection methods are used to produce bacterial, mammalian, yeast or insect cell lines that express large quantities of protein, which are then purified using standard techniques (see, e.g., Colley et al., *J. Biol. Chem.* 264:17619-17622 (1989); *Guide to Protein Purification*, in *Methods in Enzymology*, vol. 182 (Deutscher, ed., 1990)). Transformation of

eukaryotic and prokaryotic cells are performed according to standard techniques (see, e.g., Morrison, J. Bact. 132:349-351 (1977); Clark-Curtiss & Curtiss, Methods in Enzymology 101:347-362 (Wu et al., eds, 1983).

Any of the well known procedures for introducing foreign nucleotide sequences into host cells may be used. These include the use of calcium phosphate transfection, polybrene, protoplast fusion, electroporation, liposomes, microinjection, naked DNA, plasmid vectors, viral vectors, both episomal and integrative, and any of the other well known methods for introducing cloned genomic DNA, cDNA, synthetic DNA or other foreign genetic material into a host cell (see, e.g., Sambrook et al., supra). It is only necessary that the particular genetic engineering procedure used be capable of successfully introducing at least one gene into the host cell capable of expressing the protein of choice.

ii. Assays For Determining Regulation of Gene Expression

A variety of assays can be used to determine the level of gene expression regulation by the NRSF-based Zf proteins, see for example U.S. Patent No. 6,453,242. The activity of a particular NRSF-based Zf protein can be assessed using a variety of *in vitro* and *in vivo* assays, by measuring, e.g., protein or mRNA levels, product levels, enzyme activity, tumor growth; transcriptional activation or repression of a reporter gene; second messenger levels (e.g., cGMP, cAMP, IP3, DAG, Ca^{sup.2+}); cytokine and hormone production levels; and neovascularization, using, e.g., immunoassays (e.g., ELISA and immunohistochemical assays with antibodies), hybridization assays (e.g., RNase protection, northern, in situ hybridization, oligonucleotide array studies), colorimetric assays, amplification assays, enzyme activity assays, tumor growth assays, phenotypic assays, and the like.

NRSF-based Zf proteins are typically first tested for activity *in vitro* using cultured cells, e.g., 293 cells, CHO cells, VERO cells, BHK cells, HeLa cells, COS cells, and the like. Preferably, human cells are used. The NRSF-based Zf protein is often first tested using a transient expression system with a reporter gene, and then regulation of the target endogenous gene is tested in cells and in animals, both *in vivo* and *ex vivo*. The NRSF-based Zf protein can be recombinantly expressed in a cell, recombinantly expressed in cells transplanted into an animal, or recombinantly expressed in a transgenic animal, as well as administered as a protein to an animal or cell using delivery vehicles described below. The cells can be immobilized, be in

solution, be injected into an animal, or be naturally occurring in a transgenic or non-transgenic animal.

Modulation of gene expression is tested using one of the *in vitro* or *in vivo* assays described herein. Samples or assays are treated with the NRSF-based Zf protein and compared to un-treated control samples, to examine the extent of modulation. For regulation of endogenous gene expression, the NRSF-based Zf protein ideally has a K_d of 200 nM or less, more preferably 100 nM or less, more preferably 50 nM, most preferably 25 nM or less. (The effects of the NRSF-based Zf protein can be measured by examining any of the parameters described above. Any suitable gene expression, phenotypic, or physiological change can be used to assess the influence of the NRSF-based Zf protein. When the functional consequences are determined using intact cells or animals, one can also measure a variety of effects such as tumor growth, neovascularization, hormone release, transcriptional changes to both known and uncharacterized genetic markers (e.g., northern blots or oligonucleotide array studies), changes in cell metabolism such as cell growth or pH changes, and changes in intracellular second messengers such as cGMP.

Preferred assays for ZFP regulation of endogenous gene expression can be performed *in vitro*. In one *in vitro* assay format, the NRSF-based Zf protein regulation of endogenous gene expression in cultured cells is measured by examining protein production using an ELISA assay. The test sample is compared to control cells treated with an empty vector or an unrelated Zf protein that is targeted to another gene.

In another embodiment, regulation of endogenous gene expression is determined *in vitro* by measuring the level of target gene mRNA expression. The level of gene expression is measured using amplification, e.g., using RT-PCR, LCR, or hybridization assays, e.g., northern hybridization, RNase protection, dot blotting. RNase protection is used in one embodiment. The level of protein or mRNA is detected using directly or indirectly labeled detection agents, e.g., fluorescently or radioactively labeled nucleic acids, radioactively or enzymatically labeled antibodies, and the like, as described herein.

Alternatively, a reporter gene system can be devised using the target gene promoter operably linked to a reporter gene such as luciferase, green fluorescent protein, CAT, or .beta.-gal. The reporter construct is typically co-transfected into a cultured cell. After treatment with

the NRSF-based Zf protein the amount of reporter gene transcription, translation, or activity is measured according to standard techniques known to those of skill in the art.

Another example of an assay format useful for monitoring regulation of endogenous gene expression is performed *in vivo*. This assay is particularly useful for examining Zf proteins that inhibit expression of tumor promoting genes, genes involved in tumor support, such as neovascularization (e.g., VEGF), or that activate tumor suppressor genes such as p53. In this assay, cultured tumor cells expressing the NRSF-based Zf protein are injected subcutaneously into an immune compromised mouse such as an athymic mouse, an irradiated mouse, or a SCID mouse. After a suitable length of time, preferably 4-8 weeks, tumor growth is measured, e.g., by volume or by its two largest dimensions, and compared to the control. Tumors that have statistically significant reduction (using, e.g., Student's T test) are said to have inhibited growth. Alternatively, the extent of tumor neovascularization can also be measured. Immunoassays using endothelial cell specific antibodies are used to stain for vascularization of the tumor and the number of vessels in the tumor. Tumors that have a statistically significant reduction in the number of vessels (using, e.g., Student's T test) are said to have inhibited neovascularization.

Transgenic and non-transgenic animals can also be used for examining regulation of endogenous gene expression *in vivo*. Transgenic animals typically express the NRSF-based Zf protein. Alternatively, animals that transiently express the NRSF-based Zf protein, or to which the NRSF-based Zf protein has been administered in a delivery vehicle, can be used. Regulation of endogenous gene expression is tested using any one of the assays described herein.

iii. Nucleic Acids Encoding Fusion Proteins and Gene Therapy

The NRSF-based proteins of the present invention can be used to regulate gene expression in gene therapy applications in the same was as has already been described for other types of synthetic zinc finger proteins, see for example U.S. Patent No. 6,511,808, U.S. Patent No. 6,013,453, U.S. Patent No. 6,007,988, U.S. Patent No. 6,503,717, U.S. patent application 2002/0164575 A1, and U.S. patent application 2002/0160940 A1.

Conventional viral and non-viral based gene transfer methods can be used to introduce nucleic acids encoding the NRSF-based Zf protein into mammalian cells or target tissues. Such methods can be used to administer nucleic acids encoding the NRSF-based Zf proteins to cells *in vitro*. Preferably, the nucleic acids encoding the NRSF-based Zf protein s are administered for *in vivo* or *ex vivo* gene therapy uses. Non-viral vector delivery systems include DNA plasmids,

naked nucleic acid, and nucleic acid complexed with a delivery vehicle such as a liposome. Viral vector delivery systems include DNA and RNA viruses, which have either episomal or integrated genomes after delivery to the cell. For a review of gene therapy procedures, see Anderson, Science 256:808-813 (1992); Nabel & Felgner, TIBTECH 11:211-217 (1993); Mitani & Caskey, TIBTECH 11:162-166 (1993); Dillon, TIBTECH 11:167-175 (1993); Miller, Nature 357:455-460 (1992); Van Brunt, Biotechnology 6(10):1149-1154 (1988); Vigne, Restorative Neurology and Neuroscience 8:35-36 (1995); Kremer & Perricaudet, British Medical Bulletin 51(1):31-44 (1995); Haddada et al., in Current Topics in Microbiology and Immunology Doerfler and Bohm (eds) (1995); and Yu et al., Gene Therapy 1:13-26 (1994).

Methods of non-viral delivery of nucleic acids encoding the NRSF-based Zf proteins include lipofection, microinjection, biolistics, virosomes, liposomes, immunoliposomes, polycation or lipid:nucleic acid conjugates, naked DNA, artificial virions, and agent-enhanced uptake of DNA. Lipofection is described in e.g., U.S. Pat. No. 5,049,386, No. 4,946,787; and No. 4,897,355) and lipofection reagents are sold commercially (e.g., Transfectam.TM. and Lipofectin.TM.). Cationic and neutral lipids that are suitable for efficient receptor-recognition lipofection of polynucleotides include those of Felgner, WO 91/17424, WO 91/16024. Delivery can be to cells (*ex vivo* administration) or target tissues (*in vivo* administration).

The preparation of lipid:nucleic acid complexes, including targeted liposomes such as immunolipid complexes, is well known to one of skill in the art (see, e.g., Crystal, Science 270:404-410 (1995); Blaese et al., Cancer Gene Ther. 2:291-297 (1995); Behr et al., Bioconjugate Chem. 5:382-389 (1994); Remy et al., Bioconjugate Chem. 5:647-654 (1994); Gao et al., Gene Therapy 2:710-722 (1995); Ahmad et al., Cancer Res. 52:4817-4820 (1992); U.S. Pat. Nos. 4,186,183, 4,217,344, 4,235,871, 4,261,975, 4,485,054, 4,501,728, 4,774,085, 4,837,028, and 4,946,787).

The use of RNA or DNA viral based systems for the delivery of nucleic acids encoding the NRSF-based Zf proteins takes advantage of highly evolved processes for targeting a virus to specific cells in the body and trafficking the viral payload to the nucleus. Viral vectors can be administered directly to patients (*in vivo*) or they can be used to treat cells *in vitro* and the modified cells are administered to patients (*ex vivo*). Conventional viral based systems for the delivery of Zf proteins could include retroviral, lentivirus, adenoviral, adeno-associated and herpes simplex virus vectors for gene transfer. Viral vectors are currently the most efficient and

versatile method of gene transfer in target cells and tissues. Integration in the host genome is possible with the retrovirus, lentivirus, and adeno-associated virus gene transfer methods, often resulting in long term expression of the inserted transgene. Additionally, high transduction efficiencies have been observed in many different cell types and target tissues.

The tropism of a retrovirus can be altered by incorporating foreign envelope proteins, expanding the potential target population of target cells. Lentiviral vectors are retroviral vector that are able to transduce or infect non-dividing cells and typically produce high viral titers. Selection of a retroviral gene transfer system would therefore depend on the target tissue. Retroviral vectors are comprised of cis-acting long terminal repeats with packaging capacity for up to 6-10 kb of foreign sequence. The minimum cis-acting LTRs are sufficient for replication and packaging of the vectors, which are then used to integrate the therapeutic gene into the target cell to provide permanent transgene expression. Widely used retroviral vectors include those based upon murine leukemia virus (MuLV), gibbon ape leukemia virus (GaLV), Simian Immuno deficiency virus (SIV), human immuno deficiency virus (HIV), and combinations thereof (see, e.g., Buchscher et al., J. Virol. 66:2731-2739 (1992); Johann et al., J. Virol. 66:1635-1640 (1992); Sommerfelt et al., Virol. 176:58-59 (1990); Wilson et al., J. Virol. 63:2374-2378 (1989); Miller et al., J. Virol. 65:2220-2224 (1991); PCT/US94/05700).

In applications where transient expression of the NRSF-based Zf protein is preferred, adenoviral based systems are typically used. Adenoviral based vectors are capable of very high transduction efficiency in many cell types and do not require cell division. With such vectors, high titer and levels of expression have been obtained. This vector can be produced in large quantities in a relatively simple system. Adeno-associated virus ("AAV") vectors are also used to transduce cells with target nucleic acids, e.g., in the *in vitro* production of nucleic acids and peptides, and for *in vivo* and *ex vivo* gene therapy procedures (see, e.g., West et al., Virology 160:38-47 (1987); U.S. Pat. No. 4,797,368; WO 93/24641; Kotin, Human Gene Therapy 5:793-801 (1994); Muzyczka, J. Clin. Invest. 94:1351 (1994). Construction of recombinant AAV vectors are described in a number of publications, including U.S. Pat. No. 5,173,414; Tratschin et al., Mol. Cell. Biol. 5:3251-3260 (1985); Tratschin, et al., Mol. Cell. Biol. 4:2072-2081 (1984); Hermonat & Muzyczka, PNAS 81:6466-6470 (1984); and Samulski et al., J. Virol. 63:03822-3828 (1989).

In particular, at least six viral vector approaches are currently available for gene transfer in clinical trials, with retroviral vectors by far the most frequently used system. All of these viral vectors utilize approaches that involve complementation of defective vectors by genes inserted into helper cell lines to generate the transducing agent.

pLASN and MFG-S are examples are retroviral vectors that have been used in clinical trials (Dunbar et al., Blood 85:3048-305 (1995); Kohn et al., Nat. Med. 1:1017-102 (1995); Malech et al., PNAS 94:22 12133-12138 (1997)). PA317/pLASN was the first therapeutic vector used in a gene therapy trial. (Blaese et al., Science 270:475-480 (1995)). Transduction efficiencies of 50% or greater have been observed for MFG-S packaged vectors. (Ellem et al., Immunol Immunother. 44(1): 10-20 (1997); Dranoff et al., Hum. Gene Ther. 1:111-2 (1997)).

Recombinant adeno-associated virus vectors (rAAV) are a promising alternative gene delivery systems based on the defective and nonpathogenic parvovirus adeno-associated type 2 virus. All vectors are derived from a plasmid that retains only the AAV 145 bp inverted terminal repeats flanking the transgene expression cassette. Efficient gene transfer and stable transgene delivery due to integration into the genomes of the transduced cell are key features for this vector system. (Wagner et al., Lancet 351:9117 1702-3 (1998), Kearns et al., Gene Ther. 9:748-55 (1996)).

Replication-deficient recombinant adenoviral vectors (Ad) are predominantly used for colon cancer gene therapy, because they can be produced at high titer and they readily infect a number of different cell types. Most adenovirus vectors are engineered such that a transgene replaces the Ad E1a, E1b, and E3 genes; subsequently the replication defector vector is propagated in human 293 cells that supply deleted gene function in trans. Ad vectors can transduce multiply types of tissues *in vivo*, including nondividing, differentiated cells such as those found in the liver, kidney and muscle system tissues. Conventional Ad vectors have a large carrying capacity. An example of the use of an Ad vector in a clinical trial involved polynucleotide therapy for antitumor immunization with intramuscular injection (Sternan et al., Hum. Gene Ther. 7:1083-9 (1998)). Additional examples of the use of adenovirus vectors for gene transfer in clinical trials include Rosenecker et al., Infection 24:15-10 (1996); Sternan et al., Hum. Gene Ther. 9:7 1083-1089 (1998); Welsh et al., Hum. Gene Ther. 2:205-18 (1995); Alvarez et al., Hum. Gene Ther. 5:597-613 (1997); Topf et al., Gene Ther. 5:507-513 (1998); Sternan et al., Hum. Gene Ther. 7:1083-1089 (1998).

Packaging cells are used to form virus particles that are capable of infecting a host cell. Such cells include 293 cells, which package adenovirus, and psi.2 cells or PA317 cells, which package retrovirus. Viral vectors used in gene therapy are usually generated by producer cell line that packages a nucleic acid vector into a viral particle. The vectors typically contain the minimal viral sequences required for packaging and subsequent integration into a host, other viral sequences being replaced by an expression cassette for the protein to be expressed. The missing viral functions are supplied in trans by the packaging cell line. For example, AAV vectors used in gene therapy typically only possess ITR sequences from the AAV genome which are required for packaging and integration into the host genome. Viral DNA is packaged in a cell line, which contains a helper plasmid encoding the other AAV genes, namely rep and cap, but lacking ITR sequences. The cell line is also infected with adenovirus as a helper. The helper virus promotes replication of the AAV vector and expression of AAV genes from the helper plasmid. The helper plasmid is not packaged in significant amounts due to a lack of ITR sequences. Contamination with adenovirus can be reduced by, e.g., heat treatment to which adenovirus is more sensitive than AAV.

In many gene therapy applications, it is desirable that the gene therapy vector be delivered with a high degree of specificity to a particular tissue type. A viral vector is typically modified to have specificity for a given cell type by expressing a ligand as a fusion protein with a viral coat protein on the viruses outer surface. The ligand is chosen to have affinity for a receptor known to be present on the cell type of interest. For example, Han et al., PNAS 92:9747-9751 (1995), reported that Moloney murine leukemia virus can be modified to express human heregulin fused to gp70, and the recombinant virus infects certain human breast cancer cells expressing human epidermal growth factor receptor. This principle can be extended to other pairs of virus expressing a ligand fusion protein and target cell expressing a receptor. For example, filamentous phage can be engineered to display antibody fragments (e.g., FAB or Fv) having specific binding affinity for virtually any chosen cellular receptor. Although the above description applies primarily to viral vectors, the same principles can be applied to nonviral vectors. Such vectors can be engineered to contain specific uptake sequences thought to favor uptake by specific target cells.

Gene therapy vectors can be delivered *in vivo* by administration to an individual patient, typically by systemic administration (e.g., intravenous, intraperitoneal, intramuscular, subdermal,

or intracranial infusion) or topical application, as described below. Alternatively, vectors can be delivered to cells *ex vivo*, such as cells explanted from an individual patient (e.g., lymphocytes, bone marrow aspirates, tissue biopsy) or universal donor hematopoietic stem cells, followed by reimplantation of the cells into a patient, usually after selection for cells which have incorporated the vector.

Ex vivo cell transfection for diagnostics, research, or for gene therapy (e.g., via re-infusion of the transfected cells into the host organism) is well known to those of skill in the art. In a preferred embodiment, cells are isolated from the subject organism, transfected with nucleic acid (gene or cDNA), encoding the NRSF-based Zf protein, and re-infused back into the subject organism (e.g., patient). Various cell types suitable for *ex vivo* transfection are well known to those of skill in the art (see, e.g., Freshney et al., *Culture of Animal Cells, A Manual of Basic Technique* (3rd ed. 1994)) and the references cited therein for a discussion of how to isolate and culture cells from patients).

In one embodiment, stem cells are used in *ex vivo* procedures for cell transfection and gene therapy. The advantage to using stem cells is that they can be differentiated into other cell types *in vitro*, or can be introduced into a mammal (such as the donor of the cells) where they will engraft in the bone marrow. Methods for differentiating CD34+ cells *in vitro* into clinically important immune cell types using cytokines such as GM-CSF, IFN- γ , and TNF- α , are known (see Inaba et al., *J. Exp. Med.* 176:1693-1702 (1992)).

Stem cells are isolated for transduction and differentiation using known methods. For example, stem cells are isolated from bone marrow cells by panning the bone marrow cells with antibodies which bind unwanted cells, such as CD4+ and CD8+ (T cells), CD45+ (panB cells), GR-1 (granulocytes), and iad (differentiated antigen presenting cells) (see Inaba et al., *J. Exp. Med.* 176:1693-1702 (1992)).

Vectors (e.g., retroviruses, adenoviruses, liposomes, etc.) containing therapeutic the NRSF-based Zf protein nucleic acids can be also administered directly to the organism for transduction of cells *in vivo*. Alternatively, naked DNA can be administered. Administration is by any of the routes normally used for introducing a molecule into ultimate contact with blood or tissue cells. Suitable methods of administering such nucleic acids are available and well known to those of skill in the art, and, although more than one route can be used to administer a

particular composition, a particular route can often provide a more immediate and more effective reaction than another route.

Pharmaceutically acceptable carriers are determined in part by the particular composition being administered, as well as by the particular method used to administer the composition. Accordingly, there is a wide variety of suitable formulations of pharmaceutical compositions available, as described below (see, e.g., Remington's Pharmaceutical Sciences, 17th ed., 1989).

iv. Delivery Vehicles

An important factor in the administration of polypeptide compounds, such as the NRSF-based Zf proteins of the present invention, is ensuring that the polypeptide has the ability to traverse the plasma membrane of a cell, or the membrane of an intra-cellular compartment such as the nucleus. Cellular membranes are composed of lipid-protein bilayers that are freely permeable to small, nonionic lipophilic compounds and are inherently impermeable to polar compounds, macromolecules, and therapeutic or diagnostic agents. However, proteins and other compounds such as liposomes have been described, which have the ability to translocate polypeptides such as NRSF-based Zf protein across a cell membrane.

For example, "membrane translocation polypeptides" have amphiphilic or hydrophobic amino acid subsequences that have the ability to act as membrane-translocating carriers. In one embodiment, homeodomain proteins have the ability to translocate across cell membranes. The shortest internalizable peptide of a homeodomain protein, Antennapedia, was found to be the third helix of the protein, from amino acid position 43 to 58 (see, e.g., Prochiantz, *Current Opinion in Neurobiology* 6:629-634 (1996)). Another subsequence, the h (hydrophobic) domain of signal peptides, was found to have similar cell membrane translocation characteristics (see, e.g., Lin et al., *J. Biol. Chem.* 270:1 4255-14258 (1995)).

Examples of peptide sequences which can be linked to a protein, for facilitating uptake of the protein into cells, include, but are not limited to: an 11 amino acid peptide of the tat protein of HIV; a 20 residue peptide sequence which corresponds to amino acids 84-103 of the p16 protein (see Fahraeus et al., *Current Biology* 6:84 (1996)); the third helix of the 60-amino acid long homeodomain of Antennapedia (Derossi et al., *J. Biol. Chem.* 269:10444 (1994)); the h region of a signal peptide such as the Kaposi fibroblast growth factor (K-FGF) h region (Lin et al., *supra*); or the VP22 translocation domain from HSV (Elliot & O'Hare, *Cell* 88:223-233

(1997)). Other suitable chemical moieties that provide enhanced cellular uptake may also be chemically linked to the NRSF-based Zf proteins of the present invention.

Toxin molecules also have the ability to transport polypeptides across cell membranes. Often, such molecules are composed of at least two parts (called "binary toxins"): a translocation or binding domain or polypeptide and a separate toxin domain or polypeptide. Typically, the translocation domain or polypeptide binds to a cellular receptor, and then the toxin is transported into the cell. Several bacterial toxins, including *Clostridium perfringens* iota toxin, diphtheria toxin (DT), *Pseudomonas* exotoxin A (PE), pertussis toxin (PT), *Bacillus anthracis* toxin, and pertussis adenylate cyclase (CYA), have been used in attempts to deliver peptides to the cell cytosol as internal or amino-terminal fusions (Arora et al., *J. Biol. Chem.*, 268:3334-3341 (1993); Perelle et al., *Infect. Immun.*, 61:5147-5156 (1993); Stenmark et al., *J. Cell Biol.* 113:1025-1032 (1991); Donnelly et al., *PNAS* 90:3530-3534 (1993); Carbonetti et al., *Abstr. Annu. Meet. Am. Soc. Microbiol.* 95:295 (1995); Sebo et al., *Infect. Immun.* 63:3851-3857 (1995); Klimpel et al., *PNAS U.S.A.* 89:10277-10281 (1992); and Novak et al., *J. Biol. Chem.* 267:17186-17193 (1992)).

Such subsequences can be used to translocate NRSF-based Zf proteins across a cell membrane. The NRSF-based Zf proteins can be conveniently fused to or derivatized with such sequences. Typically, the translocation sequence is provided as part of a fusion protein. Optionally, a linker can be used to link the NRSF-based Zf protein and the translocation sequence. Any suitable linker can be used, e.g., a peptide linker.

The NRSF-based Zf protein can also be introduced into an animal cell, preferably a mammalian cell, via a liposomes and liposome derivatives such as immunoliposomes. The term "liposome" refers to vesicles comprised of one or more concentrically ordered lipid bilayers, which encapsulate an aqueous phase. The aqueous phase typically contains the compound to be delivered to the cell, i.e., the NRSF-based Zf protein.

The liposome fuses with the plasma membrane, thereby releasing the drug into the cytosol. Alternatively, the liposome is phagocytosed or taken up by the cell in a transport vesicle. Once in the endosome or phagosome, the liposome either degrades or fuses with the membrane of the transport vesicle and releases its contents.

In current methods of drug delivery via liposomes, the liposome ultimately becomes permeable and releases the encapsulated compound (in this case, the NRSF-based Zf protein) at

the target tissue or cell. For systemic or tissue specific delivery, this can be accomplished, for example, in a passive manner wherein the liposome bilayer degrades over time through the action of various agents in the body. Alternatively, active drug release involves using an agent to induce a permeability change in the liposome vesicle. Liposome membranes can be constructed so that they become destabilized when the environment becomes acidic near the liposome membrane (see, e.g., PNAS 84:7851 (1987); Biochemistry 28:908 (1989)). When liposomes are endocytosed by a target cell, for example, they become destabilized and release their contents. This destabilization is termed fusogenesis. Dioleoylphosphatidylethanolamine (DOPE) is the basis of many "fusogenic" systems.

Such liposomes typically comprise the NRSF-based Zf protein and a lipid component, e.g., a neutral and/or cationic lipid, optionally including a receptor-recognition molecule such as an antibody that binds to a predetermined cell surface receptor or ligand (e.g., an antigen). A variety of methods are available for preparing liposomes as described in, e.g., Szoka et al., Ann. Rev. Biophys. Bioeng. 9:467 (1980), U.S. Pat. Nos. 4,186,183, 4,217,344, 4,235,871, 4,261,975, 4,485,054, 4,501,728, 4,774,085, 4,837,028, 4,235,871, 4,261,975, 4,485,054, 4,501,728, 4,774,085, 4,837,028, 4,946,787, PCT Publication No. WO 91/017424, Deamer & Bangham, Biochim. Biophys. Acta 443:629-634 (1976); Fraley, et al., PNAS 76:3348-3352 (1979); Hope et al., Biochim. Biophys. Acta 812:55-65 (1985); Mayer et al., Biochim. Biophys. Acta 858:161-168 (1986); Williams et al., PNAS 85:242-246 (1988); Liposomes (Ostro (ed.), 1983, Chapter 1); Hope et al., Chem. Phys. Lip. 40:89 (1986); Gregoriadis, Liposome Technology (1984) and Lasic, Liposomes: from Physics to Applications (1993)). Suitable methods include, for example, sonication, extrusion, high pressure/homogenization, microfluidization, detergent dialysis, calcium-induced fusion of small liposome vesicles and ether-fusion methods, all of which are well known in the art.

In certain embodiments, it is desirable to target liposomes using targeting moieties that are specific to a particular cell type, tissue, and the like. Targeting of liposomes using a variety of targeting moieties (e.g., ligands, receptors, and monoclonal antibodies) has been previously described (see, e.g., U.S. Pat. Nos. 4,957,773 and 4,603,044).

Examples of targeting moieties include monoclonal antibodies specific to antigens associated with neoplasms, such as prostate cancer specific antigen and MAGE. Tumors can also be diagnosed by detecting gene products resulting from the activation or over-expression of

oncogenes, such as ras or c-erbB2. In addition, many tumors express antigens normally expressed by fetal tissue, such as the alphafetoprotein (AFP) and carcinoembryonic antigen (CEA). Sites of viral infection can be diagnosed using various viral antigens such as hepatitis B core and surface antigens (HBVc, HBVs) hepatitis C antigens, Epstein-Barr virus antigens, human immunodeficiency type-1 virus (HIV1) and papilloma virus antigens. Inflammation can be detected using molecules specifically recognized by surface molecules which are expressed at sites of inflammation such as integrins (e.g., VCAM-1), selectin receptors (e.g., ELAM-1) and the like.

Standard methods for coupling targeting agents to liposomes can be used. These methods generally involve incorporation into liposomes lipid components, e.g., phosphatidylethanolamine, which can be activated for attachment of targeting agents, or derivatized lipophilic compounds, such as lipid derivatized bleomycin. Antibody targeted liposomes can be constructed using, for instance, liposomes which incorporate protein A (see Renneisen et al., J. Biol. Chem., 265:16337-16342 (1990) and Leonetti et al., PNAS 87:2448-2451 (1990)).

v. Dosages

For therapeutic applications, the dose of the NRSF-based transcription factor to be administered to a patient is calculated in the same way as has already been described for other types of synthetic zinc finger proteins, see for example U.S. Patent No. 6,511,808, U.S. Patent No. 6,492,117, U.S. Patent No. 6,453,242, U.S. patent application 2002/0164575 A1, and U.S. patent application 2002/0160940 A1. In the context of the present disclosure, should be sufficient to effect a beneficial therapeutic response in the patient over time. In addition, particular dosage regimens can be useful for determining phenotypic changes in an experimental setting, e.g., in functional genomics studies, and in cell or animal models. The dose will be determined by the efficacy, specificity, and K_d of the particular NRSF-based Zf protein employed, the nuclear volume of the target cell, and the condition of the patient, as well as the body weight or surface area of the patient to be treated. The size of the dose also will be determined by the existence, nature, and extent of any adverse side-effects that accompany the administration of a particular compound or vector in a particular patient.

vi. Pharmaceutical Compositions and Administration

Appropriate pharmaceutical compositions for administration of the NRSF-based transcription factors of the present invention are determined as already described for other types of synthetic zinc finger proteins, see for example U.S. Patent No. 6,511,808, U.S. Patent No. 6,492,117, U.S. Patent No. 6,453,242, U.S. patent application 2002/0164575 A1, and U.S. patent application 2002/0160940 A1. NRSF-based Zf proteins, and expression vectors encoding NRSF-based Zf proteins, can be administered directly to the patient for modulation of gene expression and for therapeutic or prophylactic applications, for example, cancer, ischemia, diabetic retinopathy, macular degeneration, rheumatoid arthritis, psoriasis, HIV infection, sickle cell anemia, Alzheimer's disease, muscular dystrophy, neurodegenerative diseases, vascular disease, cystic fibrosis, stroke, and the like. Examples of microorganisms that can be inhibited by Zf gene therapy include pathogenic bacteria, e.g., chlamydia, rickettsial bacteria, mycobacteria, staphylococci, streptococci, pneumococci, meningococci and conococci, klebsiella, proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lyme disease bacteria; infectious fungus, e.g., Aspergillus, Candida species; protozoa such as sporozoa (e.g., Plasmodia), rhizopods (e.g., Entamoeba) and flagellates (Trypanosoma, Leishmania, Trichomonas, Giardia, etc.); viral diseases, e.g., hepatitis (A, B, or C), herpes virus (e.g., VZV, HSV-1, HSV-6, HSV-II, CMV, and EBV), HIV, Ebola, adenovirus, influenza virus, flaviviruses, echovirus, rhinovirus, coxsackie virus, comovirus, respiratory syncytial virus, mumps virus, rotavirus, measles virus, rubella virus, parvovirus, vaccinia virus, HTLV virus, dengue virus, papillomavirus, poliovirus, rabies virus, and arboviral encephalitis virus, etc.

Administration of therapeutically effective amounts is by any of the routes normally used for introducing Zf proteins into ultimate contact with the tissue to be treated. The ZFPs are administered in any suitable manner, preferably with pharmaceutically acceptable carriers. Suitable methods of administering such modulators are available and well known to those of skill in the art, and, although more than one route can be used to administer a particular composition, a particular route can often provide a more immediate and more effective reaction than another route.

Pharmaceutically acceptable carriers are determined in part by the particular composition being administered, as well as by the particular method used to administer the composition.

Accordingly, there is a wide variety of suitable formulations of pharmaceutical compositions that are available (see, e.g., Remington's Pharmaceutical Sciences, 17^{sup}.th ed. 1985)).

The ZFPs, alone or in combination with other suitable components, can be made into aerosol formulations (i.e., they can be "nebulized") to be administered via inhalation. Aerosol formulations can be placed into pressurized acceptable propellants, such as dichlorodifluoromethane, propane, nitrogen, and the like.

Formulations suitable for parenteral administration, such as, for example, by intravenous, intramuscular, intradermal, and subcutaneous routes, include aqueous and non-aqueous, isotonic sterile injection solutions, which can contain antioxidants, buffers, bacteriostats, and solutes that render the formulation isotonic with the blood of the intended recipient, and aqueous and non-aqueous sterile suspensions that can include suspending agents, solubilizers, thickening agents, stabilizers, and preservatives. The disclosed compositions can be administered, for example, by intravenous infusion, orally, topically, intraperitoneally, intravesically or intrathecally. The formulations of compounds can be presented in unit-dose or multi-dose sealed containers, such as ampules and vials. Injection solutions and suspensions can be prepared from sterile powders, granules, and tablets of the kind previously described.

vii. Regulation of Gene Expression in Plants

NRSF-based Zf proteins can be used to engineer plants for traits such as increased disease resistance, modification of structural and storage polysaccharides, flavors, proteins, and fatty acids, fruit ripening, yield, color, nutritional characteristics, improved storage capability, and the like. In particular, the engineering of crop species for enhanced oil production, e.g., the modification of the fatty acids produced in oilseeds, is of interest.

Seed oils are composed primarily of triacylglycerols (TAGs), which are glycerol esters of fatty acids. Commercial production of these vegetable oils is accounted for primarily by six major oil crops (soybean, oil palm, rapeseed, sunflower, cotton seed, and peanut.) Vegetable oils are used predominantly (90%) for human consumption as margarine, shortening, salad oils, and frying oil. The remaining 10% is used for non-food applications such as lubricants, oleochemicals, biofuels, detergents, and other industrial applications.

The desired characteristics of the oil used in each of these applications varies widely, particularly in terms of the chain length and number of double bonds present in the fatty acids making up the TAGs. These properties are manipulated by the plant in order to control

membrane fluidity and temperature sensitivity. The same properties can be controlled using NRSF-based Zf protein to produce oils with improved characteristics for food and industrial uses.

The primary fatty acids in the TAGs of oilseed crops are 16 to 18 carbons in length and contain 0 to 3 double bonds. Palmitic acid (16:0 [16 carbons: 0 double bonds]), oleic acid (18:1), linoleic acid (18:2), and linolenic acid (18:3) predominate. The number of double bonds, or degree of saturation, determines the melting temperature, reactivity, cooking performance, and health attributes of the resulting oil.

The enzyme responsible for the conversion of oleic acid (18: 1) into linoleic acid (18:2) (which is then the precursor for 18:3 formation) is .DELTA.12-oleate desaturase, also referred to as omega-6 desaturase. A block at this step in the fatty acid desaturation pathway should result in the accumulation of oleic acid at the expense of polyunsaturates.

In one embodiment NRSF-based Zf proteins are used to regulate expression of the FAD2-1 gene in soybeans. Two genes encoding microsomal DELTA.6 desaturases have been cloned recently from soybean, and are referred to as FAD2-1 and FAD2-2 (Heppard et al., *Plant Physiol.* 110:311-319 (1996)). FAD2-1 (δ -12 desaturase) appears to control the bulk of oleic acid desaturation in the soybean seed. NRSF-based Zf proteins can thus be used to modulate gene expression of FAD2-1 in plants. Specifically, NRSF-based Zf proteins can be used to inhibit expression of the FAD2-1 gene in soybean in order to increase the accumulation of oleic acid (18:1) in the oil seed. Moreover, NRSF-based Zf proteins can be used to modulate expression of any other plant gene, such as delta-9 desaturase, delta-12 desaturases from other plants, delta-15 desaturase, acetyl-CoA carboxylase, acyl-ACP-thioesterase, ADP-glucose pyrophosphorylase, starch synthase, cellulose synthase, sucrose synthase, senescence-associated genes, heavy metal chelators, fatty acid hydroperoxide lyase, polygalacturonase, EPSP synthase, plant viral genes, plant fungal pathogen genes, and plant bacterial pathogen genes.

Recombinant DNA vectors suitable for transformation of plant cells are also used to deliver protein (e.g., NRSF-based Zf proteins)-encoding nucleic acids to plant cells. Techniques for transforming a wide variety of higher plant species are well known and described in the technical and scientific literature (see, e.g., Weising et al. *Ann. Rev. Genet.* 22:421-477 (1988)). A DNA sequence coding for the desired ZFP is combined with transcriptional and translational

initiation regulatory sequences which will direct the transcription of the ZFP in the intended tissues of the transformed plant.

For example, a plant promoter fragment may be employed which will direct expression of the NRSF-based Zf protein in all tissues of a regenerated plant. Such promoters are referred to herein as "constitutive" promoters and are active under most environmental conditions and states of development or cell differentiation. Examples of constitutive promoters include the cauliflower mosaic virus (CaMV) 35 S transcription initiation region, the 1'- or 2'-promoter derived from T-DNA of *Agrobacterium tumefaciens*, and other transcription initiation regions from various plant genes known to those of skill.

Alternatively, the plant promoter may direct expression of the NRSF-based Zf protein in a specific tissue or may be otherwise under more precise environmental or developmental control. Such promoters are referred to here as "inducible" promoters. Examples of environmental conditions that may effect transcription by inducible promoters include anaerobic conditions or the presence of light.

Examples of promoters under developmental control include promoters that initiate transcription only in certain tissues, such as fruit, seeds, or flowers. For example, the use of a polygalacturonase promoter can direct expression of the ZFP in the fruit, a CHS-A (chalcone synthase A from petunia) promoter can direct expression of the ZFP in flower of a plant.

The vector comprising the ZFP sequences will typically comprise a marker gene which confers a selectable phenotype on plant cells. For example, the marker may encode biocide resistance, particularly antibiotic resistance, such as resistance to kanamycin, G4 18, bleomycin, hygromycin, or herbicide resistance, such as resistance to chlorosulfuron or Basta.

Such DNA constructs may be introduced into the genome of the desired plant host by a variety of conventional techniques. For example, the DNA construct may be introduced directly into the genomic DNA of the plant cell using techniques such as electroporation and microinjection of plant cell protoplasts, or the DNA constructs can be introduced directly to plant tissue using biolistic methods, such as DNA particle bombardment. Alternatively, the DNA constructs may be combined with suitable T-DNA flanking regions and introduced into a conventional *Agrobacterium tumefaciens* host vector. The virulence functions of the *Agrobacterium tumefaciens* host will direct the insertion of the construct and adjacent marker into the plant cell DNA when the cell is infected by the bacteria.

Microinjection techniques are known in the art and well described in the scientific and patent literature. The introduction of DNA constructs using polyethylene glycol precipitation is described in Paszkowski et al. EMBO J. 3:2717-2722 (1984). Electroporation techniques are described in Fromm et al. PNAS 82:5824 (1985). Biolistic transformation techniques are described in Klein et al. Nature 327:70-73 (1987).

Agrobacterium tumefaciens-mediated transformation techniques are well described in the scientific literature (see, e.g., Horsch et al Science 233:496-498 (1984)); and Fraley et al. PNAS 80:4803 (1983)).

Transformed plant cells which are derived by any of the above transformation techniques can be cultured to regenerate a whole plant which possesses the transformed genotype and thus the desired ZFP-controlled phenotype. Such regeneration techniques rely on manipulation of certain phytohormones in a tissue culture growth medium, typically relying on a biocide and/or herbicide marker which has been introduced together with the ZFP nucleotide sequences. Plant regeneration from cultured protoplasts is described in Evans et al., Protoplasts Isolation and Culture, Handbook of Plant Cell Culture, pp. 124-176 (1983); and Binding, Regeneration of Plants, Plant Protoplasts, pp. 21-73 (1985). Regeneration can also be obtained from plant callus, explants, organs, or parts thereof. Such regeneration techniques are described generally in Klee et al. Ann. Rev. of Plant Phys. 38:467-486 (1987).

viii. Functional Genomics Assays

NRSF-based Zf proteins also have use for assays to determine the phenotypic consequences and function of gene expression. The recent advances in analytical techniques, coupled with focussed mass sequencing efforts have created the opportunity to identify and characterize many more molecular targets than were previously available. This new information about genes and their functions will speed along basic biological understanding and present many new targets for therapeutic intervention. In some cases analytical tools have not kept pace with the generation of new data. An example is provided by recent advances in the measurement of global differential gene expression. These methods, typified by gene expression microarrays, differential cDNA cloning frequencies, subtractive hybridization and differential display methods, can very rapidly identify genes that are up or down-regulated in different tissues or in response to specific stimuli. Increasingly, such methods are being used to explore biological processes such as, transformation, tumor progression, the inflammatory response, neurological

disorders etc. One can now very easily generate long lists of differentially expressed genes that correlate with a given physiological phenomenon, but demonstrating a causative relationship between an individual differentially expressed gene and the phenomenon is difficult. Until now, simple methods for assigning function to differentially expressed genes have not kept pace with the ability to monitor differential gene expression.

Using conventional molecular approaches, over expression of a candidate gene can be accomplished by cloning a full-length cDNA, subcloning it into a mammalian expression vector and transfecting the recombinant vector into an appropriate host cell. This approach is straightforward but labor intensive, particularly when the initial candidate gene is represented by a simple expressed sequence tag (EST). Under expression of a candidate gene by "conventional" methods is yet more problematic. Antisense methods and methods that rely on targeted ribozymes are unreliable, succeeding for only a small fraction of the targets selected. Gene knockout by homologous recombination works fairly well in recombinogenic stem cells but very inefficiently in somatically derived cell lines. In either case large clones of syngeneic genomic DNA (on the order of 10 kb) should be isolated for recombination to work efficiently.

The NRSF-based Zf technology of the present invention can be used to rapidly analyze differential gene expression studies. NRSF-based Zf proteins can be readily used to up or down-regulate any endogenous target gene. Very little sequence information is required to create a gene-specific DNA binding domain. This makes the NRSF-based Zf technology ideal for analysis of long lists of poorly characterized differentially expressed genes. One can simply build a zinc finger-based DNA binding domain for each candidate gene, create chimeric up and down-regulating artificial transcription factors and test the consequence of up or down-regulation on the phenotype under study (transformation, response to a cytokine etc.) by switching the candidate genes on or off one at a time in a model system.

This specific example of using engineered ZFPs to add functional information to genomic data is merely illustrative. Any experimental situation that could benefit from the specific up or down-regulation of a gene or genes could benefit from the reliability and ease of use of NRSF-based Zf proteins.

Additionally, greater experimental control can be imparted by NRSF-based Zf proteins than can be achieved by more conventional methods. This is because the production and/or function of NRSF-based Zf proteins, like other Zf proteins, can be placed under small molecule

control. Examples of this approach are provided by the Tet-On system, the ecdysone-regulated system and a system incorporating a chimeric factor including a mutant progesterone receptor. These systems are all capable of indirectly imparting small molecule control on any endogenous gene of interest or any transgene by placing the function and/or expression of a NRSF-based Zf protein under small molecule control.

ix. Transgenic Mice

A further application of NRSF-based Zf proteins is manipulating gene expression in transgenic animals. As with cell lines, over-expression of an endogenous gene or the introduction of a heterologous gene to a transgenic animal, such as a transgenic mouse, is a fairly straightforward process. NRSF-based Zf proteins technology is an improvement in these types of methods because one can circumvent the need for generating full-length cDNA clones of the gene under study.

Likewise, as with cell-based systems, conventional down-regulation of gene expression in transgenic animals is plagued by technical difficulties. Gene knockout by homologous recombination is the method most commonly applied currently. This method requires a relatively long genomic clone of the gene to be knocked out (ca. 10 kb). Typically, a selectable marker is inserted into an exon of the gene of interest to effect the gene disruption, and a second counter-selectable marker provided outside of the region of homology to select homologous versus non-homologous recombinants. This construct is transfected into embryonic stem cells and recombinants selected in culture. Recombinant stem cells are combined with very early stage embryos generating chimeric animals. If the chimerism extends to the germline homozygous knockout animals can be isolated by back-crossing. When the technology is successfully applied, knockout animals can be generated in approximately one year. Unfortunately two common issues often prevent the successful application of the knockout technology; embryonic lethality and developmental compensation. Embryonic lethality results when the gene to be knocked out plays an essential role in development. This can manifest itself as a lack of chimerism, lack of germline transmission or the inability to generate homozygous back crosses. Genes can play significantly different physiological roles during development versus in adult animals. Therefore, embryonic lethality is not considered a rationale for dismissing a gene target as a useful target for therapeutic intervention in adults. Embryonic lethality most often simply means that the gene of interest can not be easily studied in mouse models, using conventional methods.

Developmental compensation is the substitution of a related gene product for the gene product being knocked out. Genes often exist in extensive families. Selection or induction during the course of development can in some cases trigger the substitution of one family member for another mutant member. This type of functional substitution may not be possible in the adult animal. A typical result of developmental compensation would be the lack of a phenotype in a knockout mouse when the ablation of that gene's function in an adult would otherwise cause a physiological change. This is a kind of false negative result that often confounds the interpretation of conventional knockout mouse models.

A few new methods have been developed to avoid embryonic lethality. These methods are typified by an approach using the cre recombinase and lox DNA recognition elements. The recognition elements are inserted into a gene of interest using homologous recombination (as described above) and the expression of the recombinase induced in adult mice post-development. This causes the deletion of a portion of the target gene and avoids developmental complications. The method is labor intensive and suffers from chimerism due to non-uniform induction of the recombinase.

The use of NRSF-based Zf proteins to manipulate gene expression can be restricted to adult animals using the small molecule regulated systems described in the previous section. Expression and/or function of a zinc finger-based repressor can be switched off during development and switched on at will in the adult animals. This approach relies on the addition of the NRSF-based Zf protein expressing module only; homologous recombination is not required. Because the NRSF-based repressors are trans dominant, there is no concern about germline transmission or homozygosity. These issues dramatically affect the time and labor required to go from a poorly characterized gene candidate (a cDNA or EST clone) to a mouse model. This ability can be used to rapidly identify and/or validate gene targets for therapeutic intervention, generate novel model systems and permit the analysis of complex physiological phenomena (development, hematopoiesis, transformation, neural function etc.). Chimeric targeted mice can be derived according to Hogan et al., *Manipulating the Mouse Embryo: A Laboratory Manual*, (1988); Teratocarcinomas and Embryonic Stem Cells: A Practical Approach, Robertson, ed., (1987); and Capecchi et al., *Science* 244:1288 (1989).

EXAMPLES

The following examples are provided to describe and illustrate, but not limit, the claimed invention. Those of skill in the art will readily recognize a variety of non-critical parameters that could be changed or modified to yield essentially similar results.

Example 1

Use of the Bacterial 2-Hybrid System to study the binding of NRSF to DNA

Introduction

The use of a bacterial 2-hybrid system to identify and study Cys₂His₂ Zf proteins has already been described (Joung et al., 2000, Proceedings of the National Academy of Sciences (USA) 97:7382 and US Patent Application No. 20020119498). As shown in Figure 10, in an appropriately engineered *E. coli* strain, binding of a Zf protein to a target DNA sequence of interest can trigger transcriptional activation of a reporter gene. In this strain, the target DNA sequence is positioned upstream of a weak promoter that directs low level expression of a reporter gene. Transcription of the reporter gene can be activated by expressing 2 hybrid proteins, one a fusion of the Zf protein with a fragment of the yeast Gal1 1P protein (GP-Zf) and the other a fusion between a fragment of the yeast Gal4 protein and the *E. coli* RNA polymerase alpha subunit (α -Gal4 protein). Since the yeast Gal1 1P and Gal4 protein fragments can interact with each other, GP-Zf bound to the target DNA sequence can mediate recruitment of RNA polymerase complexes that have incorporated the α -Gal4 protein thereby stimulating transcription of the reporter gene from the weak promoter (see Figure 10). This transcriptional activation is absolutely dependent upon binding of the GP-Zf hybrid protein to the target DNA sequence positioned near the weak promoter. Thus, in this type of engineered *E. coli* cell, the level of reporter gene expression provides an indirect measure how well a Zf protein occupies the target DNA sequence of interest.

In the methods of the present invention a bacterial 2-hybrid system is utilized in 2 different ways, 1) as a reporter system for assessing how well a Zf protein can bind to a target sequence and activate transcription, and 2) as a selection system for identifying Zf variants (from large randomized libraries $>10^8$ in size) that bind to a target DNA sequence. As shown in Figure 10, to use the bacterial 2-hybrid as a reporter system, requires the creation of a bacterial 2-hybrid reporter strain ("B2H reporter strain") in which a target DNA sequence is positioned upstream of a weak promoter that directs the expression of the lacZ reporter gene. Expression of

the *lacZ* gene product can be easily quantified by performing β -galactosidase assays. To use the system as a selection system, a bacterial 2-hybrid selection strain ("B2H selection strain") is created in which a target sequence is positioned upstream of a weak promoter that directs expression of 2 co-cistronically expressed selectable markers, the yeast *HIS3* gene and the bacterial *aadA* gene (use of these markers is described in detail in Joung et al., 2000). All strains (B2H reporter or B2H selection) also harbor a plasmid that expresses the α -Gal4 fusion protein. In addition, all Zf proteins introduced into either B2H reporter strains or B2H selection strains are expressed as fusions to a Gal11P fragment.

Materials & Methods

Media. Histidine-deficient medium utilized for selections has been previously described. Where required, the following antibiotics were added: carbenicillin (50 μ g/ml in liquid medium, 100 μ g/ml in solid medium), chloramphenicol (30 μ g/ml), kanamycin (30 μ g/ml). Isopropyl β -D-thiogalactoside (IPTG, to induce protein expression), 3-aminotriazole (3-AT, a *HIS3* competitive inhibitor), and streptomycin were added at various concentrations to control selection conditions.

Plasmids and strains: The α Gal4 protein expression plasmid used has been described previously by (Joung et al., 2000, Proceedings of the National Academy of Sciences (USA) 97:7382). Zinc finger proteins (ZFPs) were expressed from vectors based on the previously described pBR-GP-Z123 plasmid (Joung et al., 2000 as above). In these plasmids the inducible *lacUV5* promoter directs the expression of a Zf protein fused to a fragment of the yeast Gal11p protein. Reporter strains for both selections and in vivo transcriptional activation assays were constructed as described (Joung et al., 2000 as above). These strains contain a single copy F'-episome with the target DNA binding site positioned immediately upstream of a weak *lac*-promoter that controls the transcription of the selectable *HIS3* and *aadA* genes (in "B2H selection strains") or the *lacZ* reporter gene (in "B2H reporter strains").

Results & Conclusions

Experiments were performed to determine whether the NRSF Zf domain could be studied using the bacterial 2-hybrid system. To do this, the NRSF DNA binding domain (Zfs 1-8) was fused to a fragment of the yeast Gal11P protein to create the GP-NRSF1-8 hybrid protein. In addition, a "B2H reporter strain" was constructed that harbors a consensus NRSE as the target sequence. Plasmids encoding the GP-NRSF1-8 protein or a Gal11P fragment (as a control) were

then each introduced into the B2H reporter strain and β -galactosidase assays performed to measure lacZ expression. As shown in figure 11, the GP-NRSF1-8 fusion protein efficiently stimulates transcription of the lacZ gene nearly 7-fold compared with the Gal11P only control. This increased lacZ expression is dependent upon binding of GP-NRSF1-8 to the consensus NRSE, as replacement of this sequence with an "inactive" NRSE (to which NRSF fails to bind *in vitro* or *in vivo*) abolishes activation (see Figure 11).

In conclusion, GP-NRSF1-8 can bind to the consensus NRSE present in the B2H reporter strain, and stimulate transcription of the associated lacZ reporter gene. Thus, the bacterial 2-hybrid system provides a useful genetic method for studying DNA binding by the NRSF Zf domain.

Example 2

The NRSF Zf domain binds to the NRSE with high specificity

The bacterial 2-hybrid system was used as a genetic method to assess the specificity of DNA binding by the NRSF Zf domain. Joung and colleagues have demonstrated that only Zf proteins that bind with high affinity and specificity for their target DNA sequence can activate transcription efficiently in the bacterial 2-hybrid system (Hurt et al., 2003, manuscript in preparation). Thus, this system provides a rapid method to assess how well a given Zf protein can recognize a target site of interest. A series of "B2H reporter strains," was generated, each bearing one of the mutated NRSE sequences shown in Figure 11 as a target sequence. These mutated sites bear single or clustered double or triple base pair substitutions in positions distributed throughout the consensus NRSE. To assess the effects of these mutations on DNA binding by the wild-type NRSF Zf domain, plasmids expressing the GP-NRSF1-8 or the Gal11P control fragment were introduced into each of the B2H reporter strains and β -galactosidase activities assessed. As shown in Figure 11, many of the mutations introduced in the consensus NRSE resulted in near complete loss of transcriptional activation. This sensitivity of binding to small changes throughout the length of the NRSE strongly suggests that NRSF simultaneously contacts many of the bases within the NRSE, and binds to a sequence that spans at least 20 of the 21 bases in the NRSE with exquisite specificity. There also appeared to be a correlation between the effect of mutating particular bases and the degree to which these bases are conserved in the NRSE consensus sequence (see mutant NRSE sites in Figure 11 and the conserved bases in Figure 12). It appeared that changing more strongly conserved bases resulted in greater loss of

transcriptional activation. Since the consensus is based on functionally defined NRSE sequences, this correlation suggests that the binding of NRSF to the consensus NRSE in our 2-hybrid system accurately reflects the physiologic interaction.

Example 3

Model for binding of NRSF fingers 3-8 to the NRSE

Our existing understanding of how Zfs bind to and recognize specific DNA sites permits a limited ability to predict the DNA sequences likely to be bound by a given finger. A single Zf commonly uses residues within or adjacent to its recognition helix to make contacts with bases in the major groove of DNA. Structural information combined with results from studies of designed synthetic Zfs have together demonstrated a collection of contacts that occur between amino acids at specific positions in the recognition helix and 4 consecutive bases on a DNA strand. In addition, to a first approximation, fingers linked in tandem (particularly those connected by canonical TGEKP-type linkers) will recognize adjacent stretches of DNA (that is, large gaps do not typically occur in the sequence of recognized bases).

Using published and unpublished experience in postulating amino acid-base pair contacts, a model that plausibly matches the residues in the recognition helices of NRSF fingers 3 through 8 with the NRSE sequence was constructed. The proposed alignment (shown in Figure 12) match the typical directional "polarity" of Zfs to DNA (N-terminal to C-terminal protein "reads" 3' to 5' DNA sequence) and postulates plausible contacts along a span of 17 base pairs of DNA. It is interesting to note that this model suggests that there are no contacts to certain weakly conserved positions in the consensus NRSE. In addition, the model suggests that there are no contacts for any residues in finger 6, and because this finger is also positioned over a weakly conserved region of the consensus, it is tempting to speculate that finger 6 may not make sequence-specific DNA binding contacts with the NRSE.

Example 4

NRSF Zf 1 and Zf 2 are not required for DNA binding

Introduction

The model for the NRSF/NRSE interaction described in Example 3 does not assign a direct DNA binding role to either fingers 1 or 2. These fingers are separated from each other, and from fingers 3 through 8, by linkers longer than the 5 residue linkers present between the remaining fingers. It was therefore hypothesized that these 2 fingers may not participate directly

in DNA binding. To test this hypothesis, we used the bacterial 2-hybrid system and *in vitro* DNA-binding assays to determine whether fingers 3 through 8 from NRSF were sufficient to bind the NRSE.

Materials and Methods

Bacterial Media. Histidine-deficient medium utilized for selections has been previously described. Where required, the following antibiotics were added: carbenicillin (50 µg/ml in liquid medium, 100 µg/ml in solid medium), chloramphenicol (30 µg/ml), kanamycin (30 µg/ml). Isopropyl β-D-thiogalactoside (IPTG, to induce protein expression), 3-aminotriazole (3-AT, a HIS3 competitive inhibitor), and streptomycin were added at various concentrations to control selection conditions.

Bacterial Plasmids and Strains: The αGal4 protein expression plasmid used has been described previously by (Joung et al., 2000, Proceedings of the National Academy of Sciences (USA) 97:7382). Zinc finger proteins (ZFPs) were expressed from vectors based on the previously described pBR-GP-Z123 plasmid (Joung et al., 2000 as above). In these plasmids the inducible *lacUV5* promoter directs the expression of a Zf protein fused to a fragment of the yeast Gal11p protein. Reporter strains for both selections and *in vivo* transcriptional activation assays were constructed as described (Joung et al., 2000 as above). These strains contain a single copy F'-episome with the target DNA binding site positioned immediately upstream of a weak *lac*-promoter that controls the transcription of the selectable *HIS3* and *aadA* genes (in "B2H selection strains") or the *lacZ* reporter gene (in "B2H reporter strains").

Protein expression and purification. Maltose binding protein – zinc finger protein fusions (MBP-ZFP) were expressed from a T7 promoter (plasmid pEXP1-DEST, Invitrogen, Carlsbad, CA) in the Expressway coupled *in vitro* transcription/translation system (Invitrogen, Carlsbad, CA). Proteins were expressed according to the manufacturer's instructions at 37° C for 3.5 hours with the addition of 500uM ZnCl₂ and the omission of the post-synthesis RNase A treatment. Two to three synthesis reactions for each protein were pooled and the MBP-ZFP were batch affinity purified using amylose resin (New England Biolabs). Amylose beads were washed three times with 1ml of WB1 [15mM HEPES pH 7.8, 200 mM NaCl, 1mM EDTA, 20 uM ZnSO₄, 1mM DTT] prior to the addition of protein. Proteins were allowed to bind to beads in a total volume of 750µl while rotating for 1.5 hours at 4° C. After binding, the slurry was spun at 2 x g for 3 minutes at 4° C and unbound proteins and *in vitro* transcription/translation components

were removed from beads by pipet. Beads were subsequently washed twice with 700 μ l WB1 and twice more with 700 μ l WB2 [binding buffer from Greisman and Pabo, Science (1997) with omission of acetylated BSA and addition of 1mM DTT]. After the final centrifugation, supernatant was removed and beads were resuspended in 200 μ l elution buffer [WB2 + 40mM maltose]. Elution reactions were rotated at 22° C for 30 minutes and supernatant containing MBP-ZFP was aliquoted and frozen for storage at -80° C.

Electrophoretic Mobility Shift Assays (EMSA). Gel shift assays were performed as previously described by Greisman and Pabo, Science (1997). except that a) binding buffer contained non-acetylated bovine serum albumin (100ug/ml), b) 0.5 pM or 1 pM of the labeled DNA site was used for each binding reaction, and c) protein-DNA mixtures were incubated for 1 or 4 hours at room temperature. Results for both incubation times were comparable indicating that the binding reactions had reached equilibrium after one hour and thus we averaged the results of all of these experiments. Reactions were subjected to gel electrophoresis on Criterion 4-20% native TBE polyacrylamide gels (Bio-Rad, Hercules, CA). Gels were dried, exposed overnight to phosphorimaging screens, and quantitated using Quantity One imaging software (Bio-Rad). In order to determine dissociation constants, the % of DNA bound (θ) was plotted against the concentration of protein [P] in each binding reaction. SigmaPlot8 (Sigma) non-linear regression software was used to fit the curve plotted above according to Equation (1) in the manuscript by Elrod-Erickson and Pabo (J Biol Chem (1999) Jul 2;274(27):19281-5) and to calculate values for the Kd of each protein. The concentration of active protein was determined for each experiment by titrating dilutions of the fusion ZFP against a fixed excess amount of unlabeled target site (12.5nM) and a small amount of labeled target site (1pM). Reactions were incubated and subjected to gel electrophoresis concurrently with those used for dissociation constant determination. Active protein concentrations ($[P]_{stock}$) were determined by plotting θ vs. 1/diln. factor according to Equation (1).

$$\theta = \frac{[P]_{stock}}{diln.factor} * \frac{1}{[DNA]_i} \quad (1)$$

Binding site competition experiments were performed as done by Greisman et al.(Science, 1997) with the exception that 0.5 or 1pM of radiolabeled target site was used. Specific and non-specific dissociation constants were averaged over at least three independent experiments ($R^2 \geq 0.90$).

In Vivo Transcriptional Activation Assays (β -galactosidase assays). DNA encoding selected Gal11p-Zf protein fusions and plasmid encoding α Gal4 were co-transformed into bacterial reporter strains containing respective targeted binding sites upstream of a weak promoter driving expression of the lacZ gene. β -galactosidase assays were performed as described previously (Joung et al., PNAS 2000).

Results and Conclusions

A hybrid protein consisting of NRSF fingers 3-8 fused to the yeast Gal11P fragment (protein GP-NRSF3-8) was expressed in B2H reporter strains harboring the consensus NRSE or a mutant NRSE as the target sequence and β -galactosidase assays were performed. The results shown in Figure 11 demonstrate that GP-NRSF3-8 can bind to the consensus NRSE and activate transcription nearly as efficiently as GP-NRSF1-8. To assess the specificity of DNA binding by GP-NRSF3-8, this protein was also expressed in the series of "B2H reporter strains" harboring mutated NRSE sites bearing single or clustered double or triple base pair substitutions in positions distributed throughout the NRSE and again β -galactosidase assays were performed. The results demonstrate that, like GP-NRSF1-8, GP-NRSF3-8 binds with great specificity to a span of at least 20 base pairs of DNA sequence (see Figure 11). However, certain mutations appear to have differential effects on binding by the NRSF1-8 and NRSF3-8 domains (in particular note the effect of mutations at the 3' end of the NRSE) suggesting that fingers 1 and/or 2 may play an indirect role in influencing DNA binding specificity.

Using electrophoretic mobility shift assays, biochemical evidence that purified NRSF3-8 binds to the consensus NRSE with an affinity that approaches that of purified NRSF1-8 for the same site was also obtained. Domains of NRSF fingers 1-8 and fingers 3-8 (NRSF1-8 and NRSF3-8) were expressed and purified as fusions to the maltose-binding protein using a standard optimized procedure. A synthetic double-stranded DNA template bearing a single consensus NRSE was radioactively labeled to high specific activity. Electrophoretic mobility shift assays using purified proteins and labeled site demonstrate that both of NRSF1-8 and NRSF3-8 can bind to the consensus NRSE site *in vitro* (see Figure 13). In addition, inspection of protein titration experiments suggest that both NRSF1-8 and NRSF3-8 bind with an apparent dissociation constant in the low picomolar range with NRSF1-8 binding somewhat more tightly than NRSF3-8. Taken together, the bacterial cell-based results and biochemical analysis strongly

suggest that fingers 1 and 2 of NRSF are not required for binding to the consensus NRSE, though they may make indirect contributions to DNA binding affinity and specificity.

Example 5

Mapping NRSF-NRSE DNA interactions by targeted re-engineering of DNA binding specificity

A targeted genetic approach was used to confirm the register and positioning of NRSF fingers 3 through 8 on the NRSE predicted by the NRSF-NRSE interaction model described in Example 3. In this approach a clustered double mutation was introduced into the NRSE and then residues in the recognition helix of the finger predicted by the model to interact with the mutated bases were randomized. If the model is correct it should be possible to isolate NRSF variants from the randomized library that bind specifically to the mutated NRSE and not to the original consensus NRSE. In genetic terms, such an altered DNA binding specificity NRSF variant would be similar to an "allele-specific" suppressor of the mutation(s) in the NRSE. The successful isolation of this type of NRSF variant would provide strong genetic confirmation of the interaction(s) predicted by the model. Alternatively, if the model is inaccurate in its predictions, then for a given mutation in the NRSE it should not be possible to isolate such variant NRSF suppressors.

In a preliminary test of this approach, 2 different clustered double mutations were introduced into the consensus NRSE targeting bases predicted to be bound by NRSF finger 4 or finger 5. Two different "B2H selection strains" were then constructed each harboring one of these mutated NRSE sites as the target DNA sequence. Two randomized libraries were also constructed both based on the GP-NRSF1-8 protein. In each library, 6 residues in the recognition helix of one NRSF finger (finger 4 or finger 5) were randomized. The 6 residues randomized, positions -1, 1, 2, 3, 5, and 6 numbered relative to the helix start, are all positions that can potentially contribute to DNA binding. Cassette mutagenesis was used to construct the libraries and the codon scheme used allowed 16 possible amino acids (all except the aromatics and cysteine) encoded by 24 codons. The theoretical size of these libraries is 24^6 or approximately 2×10^8 possible members. Each of the actual libraries we constructed had greater than 10^9 independent members (a 5-fold over-sampling of the theoretical library size).

To perform selections using the bacterial 2-hybrid system, plasmids encoding members of the randomized NRSF finger 4 or finger 5 libraries were introduced into their appropriately matched selection strain (see Materials and Methods of previous Examples). In both of these

selection strains, binding of a variant GP-NRSF1-8 fusion protein to the mutant NRSE should trigger transcriptional activation of the selectable *HIS3* and *aadA* genes. These transformed cells were then plated on medium that selects for the activated expression of both the *HIS3* and *aadA* genes. For both selection experiments, colonies were obtained on the selective medium plates and then isolated and sequenced. For both selections, the variants we isolated were all very similar in their recognition helix sequences demonstrating the success of the selection in identifying variants with a common function. As shown in Figure 14A, the recognition helices of the NRSF finger 4 variants are all very similar to one another and together define a single consensus sequence with completely conserved residues at positions -1, 2, 3, and 6. Figures 19-32 show the full sequences of the selected NRSF-variants illustrated in Figure 14. Note that finger 4 variants 1, 2, and 3 (F4v1, F4v2, F4v3) have identical sequences as shown in figure 14 and 19. The recognition helices of the NRSF finger 5 variants appear to define at least 2 different consensus sequences and again within each sub-group strong conservation of residues at positions -1, 2, 3, and 6 is seen (Figure 14B). It is interesting to note that the model predicts that the arginine at position 6 of finger 5 in the wild-type NRSF protein contacts the guanine located at base position 12 of the NRSE (see Figure12). This guanine remains unchanged in the mutated NRSE site used to select the finger 5 variants and the arginine at position 6 is strongly re-selected in the finger 5 variants. Although not conclusive, this result is consistent with the idea that this arginine contacts the guanine at position 12 of the NRSE providing further support for the model presented in Example 3.

Example 6

Re-engineered NRSF variants are true altered DNA binding specificity mutants

To confirm that the NRSF variants are truly altered (as opposed to just relaxed) in their DNA binding specificity, these proteins were tested to see how well they bind to the mutant NRSE they were selected to recognize and to the original consensus NRSE. To perform these tests, "B2H reporter strains" were constructed harboring the NRSE to be tested positioned upstream of a weak test promoter that controls expression of the *lacZ* gene. Two representative candidates from each selection (indicated by blue arrows in Figures 14A and 14B), and wild-type NRSF1-8 were introduced into each reporter strain, and β -galactosidase assays were performed, as described in Example 4. The results (shown in Figures 15A and 15B) reveal that variants tested from the F4 and F5 selections bind to their appropriate target mutant NRSE but fail to bind

to the original wild-type NRSE sequence demonstrating that they are true altered DNA binding specificity mutants. Thus, residues in NRSF finger 4 interact with base positions 16 and 17 in the NRSE and residues in NRSF finger 5 interact with base positions 13 and 14 (and possibly 12) in the NRSE. This genetic result provides the first detailed information regarding the position of NRSF fingers as they engage the NRSE.

A further aim was to determine whether the altered DNA binding specificity mutants of NRSF possessed the same exquisite specificity as the original wild-type NRSF protein. A particular aim was to learn whether a single base change in the mutant NRSE would abolish binding by a re-engineered variant. To test this possibility, additional mutant NRSE sequences were generated that each differed by one base from the double mutant NRSEs used to select the NRSF finger 4 and finger 5 variants. (Because the mutant NRSEs used in the selections differ from the consensus NRSE by 2 base changes, these newer mutant NRSEs also each differ from the consensus NRSE by a single base change.). "B2H reporter strains" were constructed harboring these new mutant NRSEs and the ability of the wild-type NRSF and the selected variants to activate transcription was assessed. The β -galactosidase results shown in Figures 15A and 15B demonstrate that most of the altered DNA binding specificity NRSF variants tested possess the exquisite specificity of the original protein i.e. changing just one of the mutated bases in the NRSE abolishes binding by the variant. This result is not entirely unexpected as the bacterial 2-hybrid system selects for proteins that bind with both high affinity and specificity to their target DNA sequences.

Example 7

Targeted re-engineering of the DNA binding specificity of fingers 6, 7 and 8 of NRSF

Example 5 above describes how a targeted genetic approach was used to alter the DNA binding specificity of NRSF fingers 4 and 5. Example 6 shows that these re-engineered NRSF variants have truly altered DNA binding specificity, as opposed to having just relaxed DNA binding. The present Example extends this approach to fingers 6, 7 and 8 of NRSF.

As described in Example 5, this approach involved first introducing various clustered double mutations into the NRSE. In this case, different point mutations were introduced into the consensus NRSE targeting bases predicted to be bound by NRSF finger 6, 7 or 8. Different "B2H selection strains" were then constructed each harboring one of these mutated NRSE sites as the target DNA sequence. Randomized libraries were then constructed based on the GP-

NRSF1-8 protein. Libraries RF6, RF7, and RF8 had amino acids in NRSF fingers 6, 7 and 8 randomized, respectively. In each of these libraries, 6 residues in the recognition helix of one NRSF finger (finger 6, 7 or 8) was randomized. The 6 residues randomized, positions -1, 1, 2, 3, 5, and 6 numbered relative to the helix start, are all positions that can potentially contribute to DNA binding. Cassette mutagenesis was used to construct the libraries and the codon scheme used to construct the finger 7 library allowed 16 possible amino acids (all except the aromatics and cysteine) encoded by 24 codons. The codon scheme used to construct the finger 6 and finger 8 libraries permitted 19 possible amino acids (all except cysteine) encoded by 24 codons. The theoretical size of these libraries is 24^6 or approximately 2×10^8 possible members. Each of the actual libraries we constructed had greater than 10^9 independent members (a 5-fold over-sampling of the theoretical library size).

To perform selections using the bacterial 2-hybrid system, plasmids encoding members of the randomized RF6, RF7, and RF8 libraries were introduced into their appropriately matched selection strains (see Materials and Methods of previous Examples). In each selection strain, binding of a re-engineered variant GP-NRSF1-8 fusion protein to a mutant NRSE should trigger transcriptional activation of the selectable *HIS3* and *aadA* genes. The transformed cells were plated on medium that selects for the activated expression of both the *HIS3* and *aadA* genes. Surviving colonies able to grow on selective medium plates were isolated and sequenced. The recognition helix sequences of eight candidates are shown with their respective binding sites in Figure 33 b (finger 6) and Figure 34 b (finger 8). Note that each set of sequences defines a consensus sequence (shown in bold text at the bottom of the finger sequences) suggesting that the selections were successful. In addition, one can postulate very likely contacts (based on our existing understanding of zinc finger recognition) between amino acids found at positions -1, 2, 3, or 6 of the consensus recognition helices and specific base positions in the mutated NRSE (indicated with arrows in Figures 33 b and 34 b).

Comparisons of the amino acid sequences and DNA binding specificities of wild-type and variant NRSF fingers combined with our existing understanding of zinc finger-DNA interactions, allow us to infer likely contacts between specific amino acid positions in NRSF recognition helices with particular base positions in the NRSE. For example, wild-type NRSF finger 8 recognizes the sequence $3' \text{GAC} 5'$ with residues KNY (at the -1, 3, and 6 recognition helix positions, respectively; see Figure 34 a) whereas one of the variants selected to bind the

base 3 mutant NRSE sequence ^{3'}GAG^{5'} has the residues KNR suggesting that recognition helix position 6 in finger 8 recognizes base 3 in the NRSE. This type of arginine to guanine contact is commonly found in a number of other previously described zinc finger DNA interfaces. All of the potential NRSF-NRSE interactions deduced from NRSF finger 4, 5, 6, 7, and 8 variants isolated to date are summarized in Figure 35 b, including a contact between NRSF finger 7 and the NRSE based on preliminary data. For comparison our original predicted model of the NRSF-NRSE interaction is provided in Figure 35 a.

In conclusion, the present Example, in conjunction with Examples 5 and 6, shows that the DNA binding specificities of NRSF fingers 4, 5, 6, 7, and 8, can be altered successfully, thus underscoring the utility of the methods of the present invention.

Example 8

Selection and characterization of NRSF variants with re-engineered DNA binding specificities

Artificial transcription factors composed of 3 to 6 synthetic Zfs fused to a transcriptional regulatory domain have been shown to function in mammalian cells to alter expression of endogenous target genes. However, because there appears to be a limit to the number of fingers that can simultaneously bind to DNA, the true specificity of these proteins remains unclear. Many of them may not specify significantly more than the 10 base pairs that can be bound by a 3-finger unit. Any given 10 base pair sequence will occur approximately 3000 times just by chance in the human genome. Thus, to affect the expression of only a single gene in a mammalian cell, it will very likely be necessary to design proteins capable of targeting sequences longer than 10 base pairs. The usefulness of Zf proteins for applications in biological research and gene therapy will be substantially enhanced by being able to make proteins that have specificity for a single address in the genome.

The NRSF protein exhibits a number of characteristics that make it an attractive framework upon which to design synthetic Zf proteins capable of recognizing DNA sequences significantly greater than 10 base pairs in length. Specifically, 1) NRSF recognizes an extended DNA sequence that is at least 20 base pairs in length, 2) NRSF binds with high specificity to its target DNA sequence, and 3) individual fingers in the NRSF DNA binding domain can be re-engineered to recognize new alternative DNA sequences.

The methods of the present invention can be used to create NRSF variants that recognize novel target sequences approximately 21 base pairs in length. The affinities and specificities of

these variants can be determined *in vitro* and their abilities to regulate the expression of an endogenous mammalian gene containing the extended target sequence can then be tested.

To create NRSF variants with novel DNA binding specificities, the CSPO Zf selection strategy is employed. This strategy (illustrated in Figure 16) involves 2 stages of selection that are both performed using the bacterial 2-hybrid system: In the first stage, separate low stringency selections are performed in parallel using different libraries in which one of the finger recognition helices is randomized. To perform these low stringency selections libraries are introduced into appropriately engineered "B2H selection strains" bearing the target subsite of interest and the transformed cells are plated on selective medium. Plasmids encoding NRSF variants that confer the ability to survive on histidine-deficient medium containing 50 μ M IPTG, 10 mM 3-AT and 20 μ g/ml streptomycin are isolated and sequenced. These low stringency selections yield pools of NRSF-based proteins which are then amplified and recombined together to form a secondary library. Recombination is performed using PCR-mediated fusion of DNA fragments encoding individual finger units that preserve the fingers identified in the primary selections. For each library, approximately 200 selected (but unsequenced) recognition helices for each finger position are first amplified using finger position-specific primers and then randomly fused together and amplified to create a pool of DNA molecules encoding "shuffled" NRSF-based proteins. These molecules are then cloned into an appropriate plasmid for expression as a Gal11P-fusion protein. Each library created using this method typically contains $>10^8$ independently derived members..

In the second stage, stringent selections are then performed using this recombined library to identify optimized multi- finger proteins that bind to the final target DNA sequence of interest. The secondary library is introduced into the appropriate "B2H selection strain" bearing the full target sequence of interest and the transformants are plated on a series of histidine-deficient selective medium plates containing various concentrations of IPTG, 3-AT, and streptomycin. Candidates chosen for sequencing and subsequent analysis are picked from the most stringent selection conditions that permit growth.

This approach is somewhat analogous to the affinity maturation process used by the immune system to optimize antibodies. Initial low stringency selections identify fingers with any reasonable affinity for the DNA targets and then the secondary higher stringency selection identifies fingers that work well together to recognize the target sequence with high affinity and

specificity. In the past this method was successfully use to isolate 3 different synthetic 3-finger proteins that bind with excellent affinity and specificity for their intended DNA target sequences (J. Hurt, S. Thibodeau, A. Hirsh, C. Pabo, J.K. Joung, manuscript submitted). This strategy can also be extended to the selection of proteins with 4 or more fingers simply by increasing the number of parallel low stringency selections performed in the first stage of the procedure. This method is used herein to re-engineer the specificity of the NRSF Zf domain using 6 separate finger selections in the primary screen.

The decision of which sequences to choose as targets for our designer Zf proteins is influenced by the details of the NRSF-NRSE interface. In the present example a "framework" sequence, a partially degenerate version of the 21 base pair consensus NRSE (e.g.— 5'NNNNNCNNCNNCNGNNCNNN³') that limits the possible sequences we can choose to target, is used. Any potential target sequence that matches this framework sequence can be used. The fixed, non-degenerate bases in this framework sequence are those that are contacted by recognition helix residues from more than one finger at the NRSF-NRSE interface. This limitation stems from the fact that alteration of one of these "finger overlap" bases might require randomization of more than one finger to recognize a new base at that position and the CSPO selection strategy utilizes libraries in which recognition helix positions from only a single finger are randomized (a restriction imposed by combinatorial issues). Another reason why it is desirable to fix the base at certain positions in the framework sequence is in case the specificity of the NRSF protein simply can not be altered at some positions. Initial results suggest that at least 4 bases recognized by 2 fingers can be altered. In this Example, 2 potential target sequences in the human VEGF-A and erbB2 genes are used in selections to identify NRSF-based variants that recognize each of these 2 target sites.

In the first stage of the re-engineering procedure, 6 low stringency selections in are performed in parallel — one for each of the 6 fingers in the final protein. Six randomized libraries based on the Gal11P-NRSF1-8 hybrid protein are produced, one for each of the 6 fingers (fingers 3 through 8) that contact the NRSE sequence. Bases contacted by a given finger are altered in a NRSE and this variant used to construct a "B2H selection strain." Selections are performed using this selection strain and the appropriately matched randomized library. Selections will be performed for each of the 6 subsites located within a larger target sequence. For each selection approximately 20 candidates are sequenced.

To perform the second stage of selections secondary libraries of NRSF variants consisting of “shuffled” combinations of the fingers selected in the initial selections are assembled. These libraries are constructed using a PCR-based *in vitro* recombination protocol which ensures that fingers selected at a given position remain in the same position in the reassembled protein (e.g.—fingers selected at the F4 position all occupy the F4 position in the recombined library). Each secondary library is constructed from approximately 35 different fingers selected at each of the 6 DNA binding finger positions and thus have a theoretical complexity of 35^6 or approximately 2×10^9 proteins. To ensure oversampling of this sequence space, secondary libraries are constructed consisting of at least 10^{10} members (a library size that can reasonably be attained in *E. coli*). “Shuffled” libraries are constructed in the context of a Gal11P-NRSF1-8 hybrid protein (i.e. all proteins will also contain wild-type NRSF fingers 1 and 2). The bacterial 2-hybrid system is used to perform high stringency selections to identify candidates from the secondary libraries that bind to the desired target sequences.

For each target sequence, at least 12 independent NRSF variants that survive the selection process are sequenced and characterized. To quantify the capability of these proteins to activate transcription in the bacterial 2-hybrid system, the 12 candidates from each selection are introduced into “B2H reporter strains” bearing the appropriate extend target sequence. Expression of *lacZ* in these strains is quantified by performing β -galactosidase assays.

Example 9

In vitro characterization of selected NRSF-based proteins

The affinity and specificity of our selected NRSF variants for their extended target sequences is characterized biochemically. For each of the 2 target sequences, at least 3 different NRSF variants are expressed and purified using standard protocols. Using electrophoretic mobility shift assays, the dissociation constant and specificity ratio of each protein for its specific target DNA sequence is determined (see Methods described in Example 4). NRSF variants that bind with a variety of specificities to their intended target sequence are identified. In particular, proteins that bind with comparable affinities to the same target site but exhibit differing specificities for that sequence are identified. Proteins exhibiting these differential properties are chosen for the next stage of analysis to assess the importance of specificity (as determined *in vitro*) on the functional specificity of these synthetic Zf domains in mammalian cells.

Example 10

Evaluating the functional activity and specificity of NRSF-based proteins in mammalian cells

The function of re-engineered NRSF variants is examined in mammalian cells. Proteins with greater specificity for their target should have improved cellular function in at least 2 ways: 1) these proteins bind fewer “unintended” target sequences and therefore affect the expression of fewer non-target genes, and 2) these proteins will require lower levels of expression to bind to their intended target sequence because they do not become “diverted” to non-target DNA sequences (i.e. the concentration of protein in the cell that is free to bind the target site will be higher). NRSF variants are converted into synthetic transcription factors and their effects on gene expression both at the intended target gene (using quantitative RT-PCR) and globally on all other genes (using microarray expression profiling) are assessed. A diagram summarizing this set of experiments is depicted in Figure 17.

For each of the 2 extended target DNA sequences, 2 NRSF variants selected in the previous step are tested. Ideally, these 2 variants have approximately equivalent affinities but different specificities for their target sequence. The experiments described involve activating expression of the endogenous human VEGF-A gene, however, the protocol can be modified to target other genes for either activation or repression. To create artificial transcriptional activator proteins, a mammalian expression plasmid (based on plasmid pcDNA5, Invitrogen) in which a hybrid protein consisting of our variant NRSF Zf domains (fingers 1-8) fused to the p65 activation domain is under the control of a strong CMV promoter that can be regulated by tetracycline repressor is constructed. This hybrid protein also includes an amino-terminal SV40 nuclear localization signal and a FLAG epitope tag on the carboxyl-terminal end (a similar fusion has been previously described for synthetic 3-finger proteins). In mammalian cells engineered to express tetracycline repressor, the CMV promoter on these expression plasmids is repressed and fusion protein is produced at low levels. Addition of a tetracycline analog such as doxycycline to the medium inactivates the DNA binding capability of tetracycline repressor and thereby leads to induction of fusion protein expression. This regulated (Tet-ON) system allows fusion protein expression to be controlled by adding doxycycline to the medium.

A series of stable cell lines each expressing a synthetic activator protein based on a different NRSF variant are created. Each of these lines is generated by transfecting human embryonic kidney cells that stably express tetracycline repressor (TRex 293 cells, Invitrogen)

with linearized plasmid encoding the artificial activator fusion protein and selecting for stable integrants that are resistant to hygromycin B (resistance to this antibiotic is encoded on the pcDNA5 expression plasmid). TRex 293 cells are used because they express low levels of VEGF-A. For each synthetic activator protein, at least 10 independent stably integrated cell lines are isolated.

To assess the abilities of the NRSF variant activators to stimulate transcription of the VEGF-A gene, quantitative RT-PCR is utilized. Stable cell lines expressing the artificial activators and a control cell line with a stably integrated pcDNA5 plasmid that does not express any activator protein is grown in the presence of doxycycline (at a concentration [1 µg/mL] that will fully induce protein expression). Total RNA is isolated from each line (RNeasy, QIAgen) and used as template for first strand DNA synthesis. Quantitative RT-PCR reactions are then performed using, for example, Taqman chemistry and an ABI 7900HT Sequence Detection System machine. The amount of template in the reactions is normalized using expression of the GAPDH gene as a control. Primers and detection probes for the VEGF-A and control GAPDH genes are used. The fold-activation of a target gene by a given NRSF variant activator can be determined by comparing the transcript levels in cells expressing the synthetic activator with levels in the control cells that do not express any activator. Typically for any given synthetic activator, the 10 stable cell lines isolated that express that protein will activate the VEGF-A gene to various levels (due to variable levels of activator expression secondary to position-dependent effects). For each target sequence, 4 stable cell lines (2 cell lines for each synthetic activator targeted to that sequence) are chosen that activate VEGF-A to approximately the same level for subsequent microarray analysis.

To assess the functional specificity of the NRSF variant activators, i.e. their effects on the levels of non-target genes, global expression profiles of the stably transfected mammalian cell lines can be obtained using Affymetrix GeneChip technology. To obtain RNA samples for this DNA microarray analysis, all cell lines – including the SAMPLE lines which stably express variant NRSF activators under doxycycline control, and the global CONTROL cell line (the parent T-REx 293 line which does not express an activator) are grown in triplicate in medium containing doxycycline for 30 hours. RNA samples from each culture will be extracted (RNeasy kit, QIAgen), quantitated by UV spectrophotometry, and screened for gross degradation via agarose gel electrophoresis. Samples then undergo additional RNA analysis, biotinylated cRNA

probe synthesis, probe hybridization to the Affymetrix human U-133A GeneChip, staining, and laser confocal scanning. Primary gene expression data (i.e., raw data) is extracted from the scanned images of the U-133A chips by Affymetrix Microarray Suite software. The U-133A GeneChip contains over 22,000 probe sets representing a substantial majority of human genes referenced in Build 133 of the UniGene database. Thus, it provides an efficient surrogate for the entire human genome for assessing and comparing the functional specificity of our synthetic activators.

As noted above, to obtain data sets suitable for statistical analysis, for each of the 8 stable cell lines RNA is isolated from 3 independent cultures and microarray analysis is performed on each sample. A normalized expression measurement for each gene on each array is extracted from the raw data by means of the current best available algorithm. In the experiment, referenced below, the RMA algorithm implemented in the Affymetrix package of Bioconductor, an open source bioinformatics tool set for use in the R statistical programming environment, was used. The effect of the synthetic activators on expression levels of each gene is inferred from fold-activation (or fold-repression) of the gene, calculated as the appropriately transformed ratio of expression levels in the "sample" cell line to levels in the "control" cell line. Statistical significance of expression fold-change for each gene is determined using the CyberT software which implements a Bayesian probabilistic approach to address the problems of high inherent noise, variability which scales with expression level, and limited replicate numbers characteristic of microarray data.

From this analysis of our global expression data, a list of all genes whose expression is significantly altered at the level of transcription by the presence of a given synthetic activator is obtained. Genes in this list have their expression altered by different mechanisms and can be categorized into 5 groups: 1) genes that harbor an exact match of the target DNA sequence in their promoter, 2) genes that harbor a sequence *similar to* the target DNA sequence in their promoter, 3) genes whose expression is altered by the recombination event required to stably integrate the synthetic activator expression vector, 4) genes affected by the altered expression of genes in the previous 3 groups (indirect or downstream effects), and 5) genes whose expression is affected by the altered expression of VEGF-A. To assess how specifically the NRSF variants bind in a mammalian cell, it is desirable to identify the genes that fall into categories 1) and 2). Genes in category 3) are identified by comparing the genes affected in the 2 independent stable

cell lines created for each synthetic activator – these genes should only be affected in one of the 2 cell lines. The number of genes affected by the activation of VEGF-A expression (category 5) is minimal since VEGF-A is a secreted protein. However, genes in category 5) can also be identified by comparing the results of the experiments that use proteins to target different DNA sequences in the VEGF-A gene – genes affected by synthetic activators targeted to different sequences are likely to be those affected by upregulation of VEGF-A. One method for attempting to separate the genes in categories 1) and 2) from those in category 3) is to search the promoters of regulated genes for exact or partial matches to the target sequence. This method, though imperfect, helps to reduce the confounding effects introduced by genes in category 3). Such an analysis has been performed in a preliminary experiment with a 3-finger protein and activated genes appear to be enriched for near matches to the target sequence compared with genes whose expression is unaffected or repressed. In addition, chromatin immunoprecipitations with antibody against the Zf activator are performed to directly verify binding to particular promoters.

With the narrowed list of affected genes obtained, a simple measure of “functional specificity” can be calculated: the reciprocal of the total number of genes with statistically significant alterations in expression level (reciprocal, so that perfect functional specificity – characterizing a transactivator that effects only its target gene and no other – equals 1, and higher/closer to one implies better specificity than lower/closer to zero). Functional specificity should correlate with specificity as determined *in vitro*. These experiments provide a measure of how functionally specific the NRSF-derived synthetic activators are in their effects in mammalian cells.

The synthetic activators constructed from NRSF variants that bind to extend DNA sequences with high specificity should have much greater functional specificity in mammalian cells than analogous activators constructed from 3-finger proteins. Not surprisingly, synthetic 3-finger activator proteins can directly affect the expression of dozens of non-target genes in mammalian cells. This is seen using a TRex 293 cell line which stably expresses a previously described synthetic activator consisting of the p65 activation domain fused to a 3-zinc-finger DNA binding domain (termed VZ-573) designed to bind a sequence located 573 bp upstream of the endogenous VEGF-A transcriptional start site. Expression of the synthetic activator is controlled by a tetracycline-inducible CMV promoter. Expression of the activator mediates reproducible activation of VEGF-A at both the mRNA and protein levels as judged by quantitative RT-PCR and ELISA assay,

respectively. RNA is isolated from the cell line stably expressing the VZ-573 synthetic activator and from the parent TRex 293 cell line (both grown in the presence of tetracycline). The resulting RNA was then hybridized to an Affymetrix U133A GeneChip. As shown in Figure 18, promoter analysis of a subset of 30 most highly activated genes (all more highly activated than VEGF-A itself) compared to 30 unaffected and the 30 most highly repressed gene sets, demonstrates striking (and statistically significant, p -values < 0.0002) enrichment of exact or near matches to the intended target site of the VZ-573 within 2500 bases of the transcription start point (see Figure 18). This suggests that the expression levels of potentially dozens of genes are directly affected by the 3-finger VZ-573 synthetic activator protein.

The invention is further described by the following numbered paragraphs:

1. A synthetic monomeric, dimeric or multimeric NRSF-based zinc-finger polypeptide comprising at least 4 zinc-fingers, wherein the synthetic NRSF-based zinc-finger polypeptide has at least one amino acid residue in at least one zinc finger that differs in sequence from the wild-type NRSF protein.
2. The synthetic NRSF-based zinc-finger polypeptide of paragraph 1, wherein the synthetic NRSF-based zinc-finger polypeptide binds to a target sequence of interest.
3. The synthetic NRSF-based zinc-finger polypeptide of paragraph 2, wherein the synthetic NRSF-based zinc-finger polypeptide does not bind to the NRSE consensus sequence.
4. The synthetic NRSF-based zinc-finger polypeptide of paragraph 2, wherein the synthetic NRSF-based zinc-finger polypeptide comprises a transcriptional regulatory domain.
5. The synthetic NRSF-based zinc-finger polypeptide of paragraph 2, wherein the synthetic NRSF-based zinc-finger polypeptide comprises a transcriptional activation domain.
6. The synthetic NRSF-based zinc-finger polypeptide of paragraph 2, wherein the synthetic NRSF-based zinc-finger polypeptide comprises a transcriptional repression or silencer domain.
7. The synthetic NRSF-based zinc-finger polypeptide of paragraph 6, wherein the transcriptional repression or silencer domain is either or both of the C-terminal and N-terminal transcriptional repression domains of the WT NRSF protein.
8. A method of regulating the transcription of a gene of interest, comprising expressing a nucleotide encoding a synthetic NRSF-based zinc-finger polypeptide according to paragraph 4, 5, 6, or 7 and contacting a DNA target sequence in a gene of interest with said expressed synthetic NRSF-based zinc-finger polypeptide.

9. A method of producing long-term silencing of a gene of interest, comprising transiently expressing a nucleotide encoding a synthetic NRSF-based zinc-finger polypeptide according to paragraph 7, and transiently contacting a DNA target sequence in a gene of interest with said expressed synthetic NRSF-based zinc-finger polypeptide.
10. A method of selecting a synthetic NRSF-based zinc-finger polypeptide that has at least 4 zinc fingers and that binds to a DNA target sequence of interest, comprising,
 - a) obtaining nucleic acid libraries encoding NRSF-based zinc finger polypeptides wherein the encoded NRSF-based zinc finger polypeptides have at least one randomized amino acid position within at least one zinc finger,
 - b) expressing said nucleic acid libraries in a polypeptide expression system to produce said NRSF-based zinc finger polypeptides,
 - c) incubating said NRSF-based zinc finger polypeptides with said DNA target sequence of interest under conditions sufficient to form binding complexes, and
 - d) selecting said NRSF-based zinc finger polypeptides that bind to said DNA target sequence of interest.
11. The method according to paragraph 10, wherein said nucleic acid libraries encoding NRSF-based zinc finger polypeptides are expressed in a phage display polypeptide expression system.
12. The method according to paragraph 10, wherein said nucleic acid libraries encoding NRSF-based zinc finger polypeptides are expressed in a eukaryotic or prokaryotic polypeptide expression system.
13. The method according to paragraph 10, wherein said nucleic acid libraries encoding NRSF-based zinc finger polypeptides are in a bacterial polypeptide expression system.
14. A method of selecting a synthetic NRSF-based zinc-finger polypeptide comprising at least 4 zinc fingers, wherein said fingers bind to a DNA target sequence of interest, comprising the steps of:
 - a) providing at least 4 different primary nucleic acid libraries encoding NRSF-based polypeptides, wherein each primary nucleic acid library encodes polypeptides having at least 3 zinc fingers with nucleotide sequences identical to those of WT NRSF and one variable zinc finger having at least one randomized amino acid residue, and wherein each

of the at least 4 different primary nucleic acid libraries encodes an NRSF-based polypeptide having a different variable zinc finger position,

- b) incubating the NRSF-based polypeptides encoded by each of said primary nucleic acid libraries with said DNA target sequence of interest under conditions sufficient to form binding complexes,
 - c) isolating pools comprising nucleic acid sequences encoding NRSF-based polypeptides, wherein said NRSF-based polypeptides comprise said binding complexes,
 - d) recombining said pools to produce a secondary nucleic acid library,
 - e) incubating said secondary nucleic acid library with said DNA target sequence of interest under conditions sufficient to form high-affinity binding complexes, and
 - f) isolating nucleic acid sequences encoding NRSF-based polypeptides, wherein said NRSF-based polypeptides comprise said high-affinity binding complexes, and wherein said NRSF-based polypeptides bind with high affinity to said DNA target sequence of interest.
15. A nucleic acid library encoding NRSF-based polypeptides comprising at least 4 zinc fingers, wherein at least three of said zinc fingers have amino acid sequences identical to those of WT NRSF; and wherein one variable zinc finger has at least one randomized amino acid residue.
16. A nucleic acid library encoding NRSF-based polypeptides according to paragraph 15, wherein the variable zinc finger corresponds to one of zinc fingers 3 to 8 of the WT NRSF protein.
17. A nucleic acid library encoding NRSF-based polypeptides according to paragraph 15, wherein three zinc fingers have amino acid sequences identical to those of the WT NRSF protein.
18. A nucleic acid library encoding NRSF-based polypeptides according to paragraph 15, wherein six amino acid residues in the variable zinc finger are randomized.
19. A nucleic acid library encoding NRSF-based polypeptides according to paragraph 18, wherein amino acid positions -1, +1, 2, 3, 5, and 6 are randomized.
20. A group of at least four nucleic acid libraries encoding NRSF-based zinc finger polypeptides, wherein each primary nucleic acid library encodes a NRSF-based polypeptide having at least 3 zinc fingers with an amino acid sequence identical to that of the WT NRSF protein and one

- variable zinc finger, and wherein each of the at least 4 different primary nucleic acid libraries encodes an NRSF-based polypeptide having a different variable zinc finger position.
21. A DNA target sequence of interest to be used in the selection of a synthetic NRSF-based zinc finger polypeptide, wherein the target sequence comprises 10 to 24 base pairs.
 22. A DNA target sequence of interest to be used in the selection of a synthetic NRSF-based zinc finger polypeptide, wherein the target sequence can be described by the consensus nucleotide sequence 5'NNNNN(C/G)NNCNNCNGNNCNNN3'.

* * *

Having thus described in detail preferred embodiments of the present invention, it is to be understood that the invention defined by the above paragraphs is not to be limited to particular details set forth in the above description as many apparent variations thereof are possible without departing from the spirit or scope of the present invention.

ABSTRACT

The present invention relates to synthetic zinc finger proteins that are selected for binding to a DNA target sequence of interest. The synthetic zinc finger proteins of the present invention are based on the sequence of the naturally occurring transcription factor, NRSF, and are capable of binding extended DNA target sequences with high specificity.

Figure 1

Human NRSF Amino Acid Sequence

MATQVMGSSGGGLFTSSGNIOMALPNDMYDLHDLSKAELAAP
QLIMLANVALTGEVNGSCCDYL VGEERQMAELMPVGDNNFSDSEEGEGLSEADIKGE
PHGLENMELRSLELSVVEPQPVFEASGAPDIYSSNKDLPPETPGAEDKGKSSKTKPFR
CKPCQYEAEESEQFVHHIRVHSAKKFFVEESAEEKAKARESGSSTAEEGDFSKGPIRC
DRCGYNTNRYDHYTAHLKHHTRAGDNERVYKCICTYTTVSEYHWRKHLRNHFRKVVY
TCGKCNVFSDRKNVYVQHVRTHGERPYKCELCYSSSQKTHLTRHMRTHSGEKPFC
DQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCP
VCDYAASKKCNLYHFKSKHPTCPNKTMDVSKVLKTKKREADLPDNITNEKTEIEQ
TKIKGDVAGKKNEKSVKAEKRDVSKKKPSNNVSVIQVTRTRKSVTEVKEMDVHTGS
NSEKFSKTKSKRKLEVDLSLHGPVNDEESTKKKKVESKSKNNSQEVPGDSKVE
ENKKQNTCMKKSTKKTLKNKSSKSKPPQKEPVEKGAQMPPQMGPAPEAVQKG
PVQVELPPPMHAQMEGAQIRPAPDEPVQMEVVEQEGPAQKELLPPVEPAQMVGAQIVL
AHMELPPPMETAQTEVAQMGPAPMEPAQMEVAQVESAPMQVQKEPVQMELSPPMEVV
QKEPVQIELSPPMEEVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVVQKEPAQREPP
PPREPPLHMEPISKKPLRKDKKEKSNMQSERARKEQVLEIENGLVPVKDSWLLKESVS
TEDLSPPSPPLPKENLREEASGDQKLLNTGEGNKEAPLQKVGAEEADESLPGLAANTN
ESTHSSSQNLNTPEGETLNGKHQTDIVCEMKMDTDQNTRENLTGINSTVEEPVSP
MLPPSAVEEREAVSKTALASPPATMAANESQEIDEDEGHSHEGSDLSDNMSEGSDDS
GLHGARPVPQESSRKNKEALA VKAAKGDFVCFCDRSFRKKGKDYSKHLNRHLVNVYY
LEEAQQQE

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

॥ श्रीगणेशाय नमः ॥

Figure 3

Mouse NRSF Amino Acid Sequence

MATQVMGQSSGGSLFNNSANMGMALTNMYDLHELKAEALAAP
QLIMLANVALTGEASGCCDYLVGEERQMAELMPVGDNHFSESEGEGLSEADLKGLE
NMELGSLELSAVEPQPVFEASAAPEIYSANKDPAPETPVAEDKCRSSKAKPFRCKPCQ
YEAEESEQVHHIRHS AKKFFVEESA EKQAKAWESGSSPAEEGEFSKGP RCDRCGY
NTNRYDHYMAHLKHHLRAGENERIYKCICTYTTVSEYHWRKHLRNHFPRKVVYTC SKC
NYFSDRKNNYVQHVRTHTGERPYKCELCPYSSSQKTHLTRHMRTHS GEKPFKCDQCNY
VASNQHEVTRHARQVHNGPKPLNCPHC DYKTADRSNEFKKHVELHVNPRQFNCPVCDYA
ASKKCNLQYHFKSKHPTCPSKRMDVSKVKLKKTKKREADLLNNAV SNEKMENEQTKTK
GDVSGKKNEKPVKAVGKDA SKEKKPGSSVSVVQVSTRIRKSAVA AETKAAEVKHTDGGQ
TGNNPEKPC KAKKNKRKKDAEAHPSEEPVNEG PVTKKKKKSECKSKISTNVPKGGGRA
EERPGVKKQASLKKGTNKT PPKTKTSKKGK LAPKGMGQTEPSSGALAQVGVSPDPA
LQAEVGTGSSSQTELPSPMDIAKSEPAQMEVSLTGPPPV EPAQMEPSPAKPPQVEAP
TYPQPPQRGPAPPTGPAPPTGPAPPTGLAEMEPSPT EPSQKEPPPSMEPPCP
EELPQAEPPPMEDCQKELPSPVEPAQIEVAQTAPTQVQEPPPVSEPPRVKPTKRSSL
RKDRAEKELSLSEMARQEQVLMGVGLVPVRDSKLLKGNKSAQDPPAPPSPPKGNR
EETPKDQEMVSDGEGTVFPLKKGPEEAGESP AELAA LKESARVSSSEQNSAMPEGG
ASHSKCQTGSSGLCDVDTEQKTDTPMKD SAAEPVSPPTPTVDRDAGSPA VVASPPIT
LAENESQEIDEDEGIHSHDGS DLSDNMSEGDDSLGHGARTPPEATSKNGKAGLAGK
VTEGEFVCIFCDRSFRKEKDYSKHLNRHLVNVVFLEEA AEEQEERE EEEQE

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

THE UNIVERSITY OF CHICAGO

1

Figure 5

Rat NRSF Amino Acid Sequence

MATQVMGQSSGGSLFNNSGNMGMALPNDMYDLHDL SKAELAAP
QLIMLANVALTGEVNGSCCDYL VGEERQMAELMPVGDNHFSDSEGELEESAELKGD
SGLDNMELRSLSVVEPQVFEASAPEVYSSNKDPAEPVAEDKCKNLKAKPFR
KPCQYEAEEQFVHHIRVHSAKFFVEESAQAKARESGASPEEGEFSGKGP
RCGYNTNR YDHYTAHLKHHLRAGDNERVYKCICTYTTVSEYHWRKHLRNHFRKVVY
CSKCNFYFSTEKNYVQHVRTHTGERTPYKCELCYSSSQKTHLTRHMRTHSGE
QCNVYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQ
CDYAAASKKCNLYHFKSKHPTCPSKTMDSVKVKKTKRREADLHRDAAAAATEQ
EQA KTKGVDA SARSRERPVKGVGKDVPEKPKPCSNASVQVTRTRKSAVETKAA
HTDGGTGNNAEKSKAKKSKRKMDAEAHPSVEPVTEGPTKKKKTESKPKTSGEV
SRVEDRKADKQASIKKGKKTALKTAKKGSKLAPKWWGHTEPSSEMAQGESPV
PALTQAVVTPSGSTQTELSPPMDIAQTEPAQMDVSQTGPPQVQRPLPVEPAQLE
QEPPQVEPPACVEPPPPVEPPCPMEPAEMEPSPPEPSQVEPPPHLEPPLPME
LPPVEDCCQKELPPVEHAQTKVAQTGPTQVAVQEEPLFCLRATSSQANQKVISPK
KEKLSVLESEMARQEQVLEVLVPRVDSQLLKASKAPDLAPPSPLPKGHLRREET
KDQEMFSDGEGNKVSPLEKGGTEEGESRAELAAPMESTALSSESSNAPDGETLH
ECQADSTAVCEMEVDTEQKTDRLVPLKDSAVEPVSNPRVDPEAAAPVVASPPIT
ESQEI DEGEIHSHDGS DLSDNMSEGSDDSLGHGARPAPQATSKSGKEGLAVKVTEG
EFVCIFCDRSFRKEKDYSKHLNRHLVNVYFLEAAEEQ

Figure 6

Rat NRSF cDNA Sequence

atggccaccc aggtgatggg gcagctctt gggagggagaa gtctttta caacagtgcc 61 aacatgggca tggccttacc caacgacatg tatgacttgc acgactctc
gaaagcttga 121 ctggggggcac ctacgctcat tatgttagcc aacgttggccc tgaatggggg agtggaaaggc 181 agctgtctgtg attaccctgggt tgggtgaagg
agacagatgg ccggagttgat gctctgttga 241 gacaaacct tttagatag cgaaggagaa gggccttgggg agtgggcttga actaaagggt 301 gaccccaagg
ggcttgggaca catgggaactg agagttttgg agctaaaggt tgaagagccc 361 cagcccgtat ttgaagcattc agctggccca gaaagtgtaca gcttgatata agtaccggcc
421 cctgaagcac cgttggcggga ggaacaatgc aagaattga agggcaaac cttcgtttgt 481 aagccatggc agtatgaagc gggagcttga gacagtttgc
tatcatcat cctgggttcc 541 agtctaaaga agttttgt ggaagagaggt gcagagagagc aagccaaagc caggggaaatc 601 gggggcttccc cgtctgaagg
ggggcagggtc ccatcggctg tgaatcgtgt 661 ggtactaata ccaaccgggt tgaatctac agggcacacc tgaagcacc caatgaaggcc 721
gggggataac agctgtgtca caagttatc atttgcagt acagtcaggt cagcggatata 781 cacttggcgg aacaccttga gaaaccttt cccagggaaa tctacatgt
tagcaaatgc 841 aactattt cgcagagaaa aataattat gttcaacag ttgaaacta cacaggggaa 901 cggccctata atgtgaaat gtttcttacc tcaagtctc
agaaagacta ttaactga 961 cacatcgtta ctactcagg tgaagagcca tttaattgt atcagtgca ttatgttggc 1021 ttaatcagg acgaagtgtac
ccgacaggc agacaggttc acaaggggcc taaccttt 1081 aattggctc actgttgata caaacaggcc gataggaggc acttaaggaa gtcagtgagg 1141
ctgaagctca acctcggca gttcaactgc cccgtgtgtg actacgggc ctcagaagag 1201 tgaatactgc agtaccatt caagttccagg caccctacct gcccacgcaa
gacagtggtac 1261 gtttccaaag tgaagcttga gaaacccagg agggggggggc ctgacttga ccttggcggc 1321 gctggccggcc cacttgaagc gacgggacaca
gagcagggc gaaacacagggg ggttggagggc 1381 tctggggagg aagtgtagag gtccttga ggcgttga gaaagtgtcc aagagagag 1441 aagccctgt
gcaatgctc tttgtgtcag gtaactacc gaaactcggaa atcagcgggt 1501 gtagactaaag cagcggggggg aaaaacaca gtttgaacaga caggaaacaa
cggcgaagaa 1561 tctctaaag ctgaagagag caaaagggag atggagggcgg agggccatcc ctgggtcggg 1621 cctgtgactg aggggacccgt gacaaagag
aaaaagagc agagcacaac caagaccagg 1681 ggggaaagtgc ctgaagggcag cagagggggg gtagaggggg cgggacaaaa gcaaaagtgt 1741 tcaataga
aaggcggggaa gaaagcggct ctcaagacta agacagctaa aaaaaggcag 1801 aactgtctc gaaagggggt gggggcacaca gaaccttct cgggagagtc
tcaagggagg 1861 gaggctcag ttctgtct cactcaggcc cactcaggcc 1921 gaggctctt ctcccatggg tatgtctcag
acagggcctg cccagtgga cgtttccag 1981 acagggggcc ctaggggtgc ctaggggtgc 2101 gaggctccat gttcagtgga gctgtgtgag atgggaaacct cccctccat gggggcttcc
agcttcccca gtttggggc cacttcca ttgggggct cggcttcca tgggggttgc 2221 ctggctcttgg tggagtgatgg tcaagagggg ctggctcttgc
2161 caggttggagc cacttcca ttgggggct cggcttcca tgggggttgc 2221 ctggctcttgg tggagtgatgg tcaagagggg ctggctcttgc
tgggggtatgc tcaagactaa 2281 gttgtcaga caggttctac tcaagggggg gctgttgaag agggagccct ttcttctc 2341 cggagccact caagttcaagc
taaccagaa gttatctccc cgaagagccg tcccaaggag 2401 aggttggagc gcttgaagtg gtagggggag caggttgaagc ttctattga gtttggctta 2461
gttctgttca gtagagcca gcttctga gtcagcaga gtcacccagg 2521 caacgggtac cacttccaaa ggggacactt agaaagggag
agacacccaa gggaccaa 2581 atgttctg acggggagag aataaaga tccctcttc agaaagggag aacagagagaa 2641 gcttgggtga gtcagagtg
gctgggtgt cccatgggat ctaccagtc ttatctct 2701 gaacaaagt caaatgcacc agatgttga acattacaa gtcaggttca ggggttactc 2761
acttgggt gttgaatgtga agttgacact agagcagaa gacagcgtgt cctctga 2821 gtagcagag tagaaccagt gttacctt aaaccaaagg tggcccttga
agcagcggca 2881 ccagctgttgg tgggttccc tctatcact ttggcgggt ctacagggaa tgaatggat 2941 gaaggcattc acagccatga tggagagtgac
ctggggcagca acatgtctga gggggagtgat 3001 gactcaggac tggatggggc tggggcagca ccacagggag ctacgtcaaa aagttgggaaag 3061 gaagggttgg
ctgtcaaaat aacttggggg gaggtttttt gttttttt tgaatgtct 3121 tttagaaagg aaaaagacta tagcaaacac ctcaatggcc atttgggtta tttgtacttc 3181
cttgaagagag cagcttgaagg gtaggggtgt acttgggtgt ccttgggggt aagcggcagag 3241 cgaatttcta a

Figure 7

Xenopus NRSF partial Amino Acid Sequence

MATQMVNQSTGNSLFCSTYSTYNSISLDNDMYGLHDLKADMAAPR
LIMLANVALTGELSSGCCDYTPGERQMAELTTVNDNSFSDSEGRLEDSPSMDIQSH
NFIMEMEPAECSKEGTSENDGTLISNTLEVEVQDKRTPSPIDDKYKCVKSKPFCRCKP
CQYKAEEEEFVHHIKIHS AKIYVDNDSNKKAAQNEADSSISESDVSKGPIQCDCRCG
YNTNRFDHYL AHLKHHNKAGENERVYKCTICTYTTVSEYHWKKHLRNHYPRILYTCSQ
CSYFSDRKNNYIQHIRTHTGERPYQCILCPYSSSQKTHLTRHMRTHSGEKPFCCEQCS
YVASNQHEVTRHARQVHNGPKPLTCPHCDYKTADRSNFKKHVELHVNPRQFLCPVCDY
AASKKCNLQYHIKSRHSGCTNITMDVSKVKLR TKKGDIGVADV DANKQTENGNIDKS
VEETVKA EKRESCGKAKKSIVNLVDGQVAKRRRLSSTQKKIKTSDARPEKILDKSRKS
SCVKRKSDLLENSNDTQTSTV

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

Figure 8

Xenopus NRSF partial cDNA sequence

```

1  ggcacgagca  gtcggttgag  acgocgattt  gagaacogtg  gacagttctt  gaatttggga
61  gggagaatgg  cgcaagtgtg  cgaccggatt  ccgaaaaagt  tataacatg  gccactcaaa
121  tggtaacca  gtctacaggt  aacagcttgt  tctgtaccag  cacttactcc  aatatctcat
181  tggacaatga  catgtatggg  ttgcatgacc  ttccaaaagc  tgatatggca  gccctcogat
241  tgataatgct  agcaaatgtg  gctctgactg  gcgaactcag  tagtggttgc  tgtgattaca
301  cgccagaagg  agaaaggcaa  atggcagaac  taacaactgt  aaatgacaac  agcttctcag
361  atagtggagg  ggataggttg  gaagattcac  ccagcatgga  tattcagtca  cacaatttta
421  taatggagat  ggagccagct  gaatgttcaa  agaagggaac  gtctgaaaaa  gatggaaactc
481  tactctctaa  tacacttgag  gtggaggttc  aaaaggataa  aaggacaccc  agcccaacag
541  atgacaaaata  caaatgtgtg  aaagcaaac  catttcggtg  caaaccttgt  cagtacaaaag
601  cagagtctga  agaagaattt  gttcatcaca  ttaagattca  cagcgtctaa  atatatgttg
661  ataatgactc  aaataaaaaa  gcgagggtta  atgaggcaga  ttctagcata  tcggagggaat
721  ctgatgtctc  caaaggacct  attcagtggt  acaggtgttg  atacaataca  aatcgttttg
781  atcactatct  ggctcattta  aagcatcaca  acaaagctgg  agaaaaatgaa  agagtataca
841  aatgtacaat  atgtacttat  actacagtca  gtgaatatca  ctggaagaaa  catctacgta
901  accattatcc  aaggatactc  tatacatgct  cacaatgttc  ctattttct  gataggaaaa
961  ataattatat  ccagcatata  agaacacata  caggagaacg  accatatcag  tgtattctat
1021  gtccctactc  aagctcacag  aaaaccact  tgaccaggca  catgcgaact  cattcaggtg
1081  agaagccttt  caaatgtgag  cagtgtagtt  atgttgcatc  caatcagcat  gaagttacac
1141  gtcatgcaag  acaggttcac  aatggaccac  aaccattaac  ttgcoctcat  tgtgactaca
1201  aaactgcaga  tcgcagtaat  ttcaagaagc  atgtagagtt  acatgttaat  cctcgacagt
1261  ttctatgccc  tgtttgtgac  tatgtgctt  ccaaaaagt  taacttgcaa  tatcatataa
1321  aatccagaca  ctccaggatgc  acaaatatca  caatggatgt  ttccaaagta  aaactgagga
1381  caaagaaagg  agacatagga  gttgcagatg  ttgatgcaaa  taagcaaaaa  gagaatggaa
1441  atataataga  taaatctgtg  gaagagaccg  ttaaagcaga  gaaaaggaaa  agctgtggga
1501  aagctaaaaa  aagtattgtt  aatttagttg  atggccaggt  tgcaaaaaaa  aggcgcttgt
1561  catctactca  gaaaaaaatt  aaaacttcag  acgcaaggcc  tgaaaaagtt  ttagataaat
1621  ccgtaagtc  tagttgtgtg  aaagaaaaat  ctgatttatt  agaaaaattt  aatgatcccc
1681  aaacaagcac  tgtg

```

Figure 9



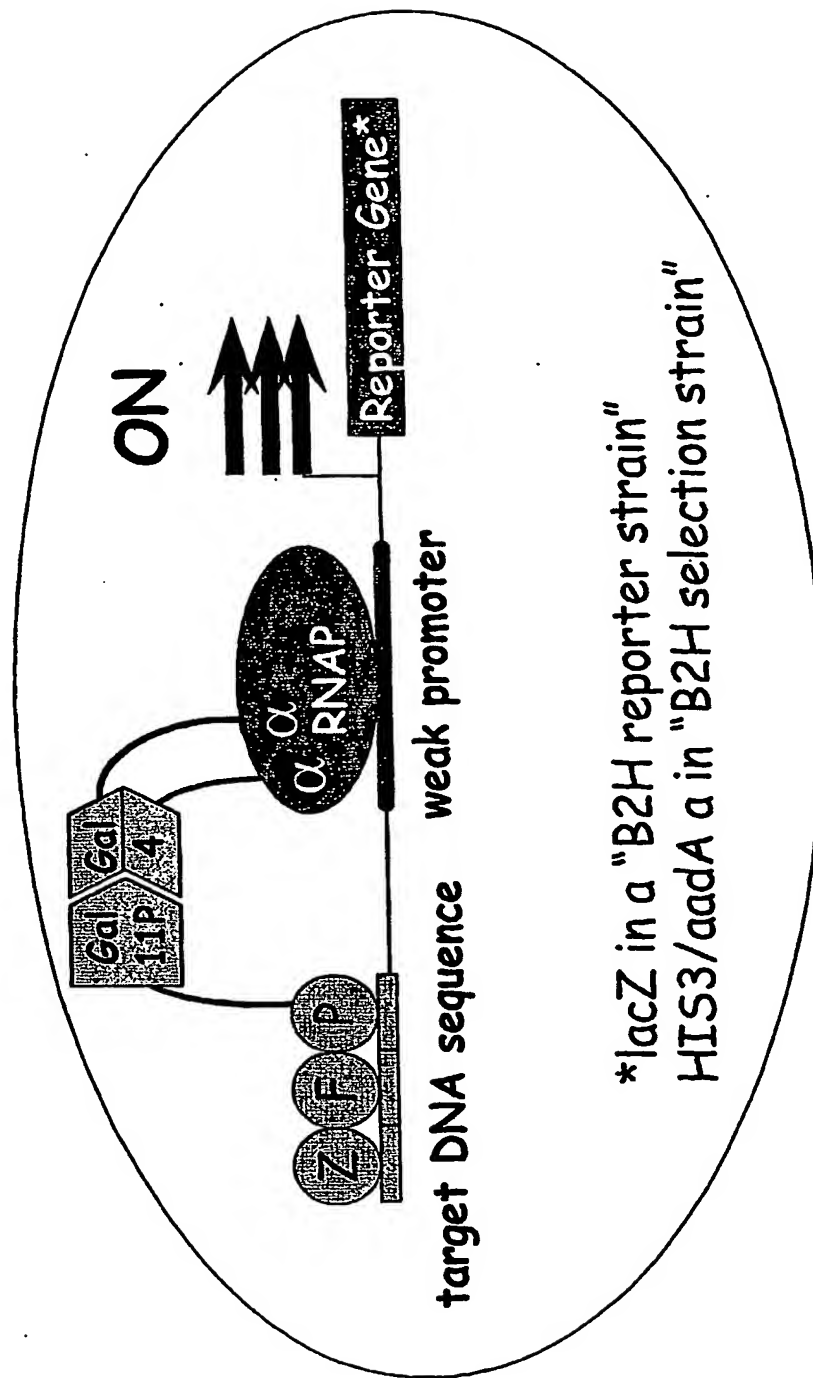
Ziff268

RPYACPVESCDRRFRRSDELTRHIRI-HTGQKP
FQCRI--CMRNFRRSDHLTTHIRT-HTGEKP
FACDI--CGRKFARSDEKRHTKI-HLRQKD

NRSF

KPFRCKP--CQYEAES~~EEQ~~FVHHIRV-HSAKKFFVEESA~~EKQ~~AKARESGSSTAEEGDFSKGP
IRCDR--CGYNTNR~~YDHY~~TAHLKH-HTRAGDNERV
YKCI--CTYTTVSEY~~HWRK~~HLRN-HFPRKV
YTCGK--CNYFS~~DRKNNY~~VQHVRT-HTGERP
YKCEL--CPYSS~~SQKTHL~~TRHMRT-HSGEKP
FKCDQ--CSYVAS~~NQHEVTR~~HARQVHNGPKP
LNC~~PH~~--CDYK~~TADRSN~~FKKHVEL-HVNP~~RQ~~
FNCPV--CDYAA~~SKKCN~~LQYHFKSKHPTCPN

Figure 10



*lacZ in a "B2H reporter strain"
HIS3/aadA a in "B2H selection strain"

Figure 11

Target DNA Sequences (NRSEs)

Consensus 3'-CCGCGACAGGCACCACGACTT-5'

Inactive 3'-TGACTCCAGGCTCCTCGACGA-5'

Mutants

3'-CGCCGACAGGCACCACGACTT-5'

3'-CCGCGTCAGGCACCACGACTT-5'

3'-CCGCCCTCAGGCACCACGACTT-5'

3'-CCGCGACACGCACCACGACTT-5'

3'-CCGCGACTCGCACCACGACTT-5'

3'-CCGCGACAGGGTCCACGACTT-5'

3'-CCGCGACAGGGTGCACGACTT-5'

3'-CCGCGACAGGCACGTCGACTT-5'

3'-CCGCGACAGGCACCACCTCTT-5'

3'-CCGCGACAGGCACCACGATAA-5'

3'-CCGCGACAGGCACCACGCAA-5'

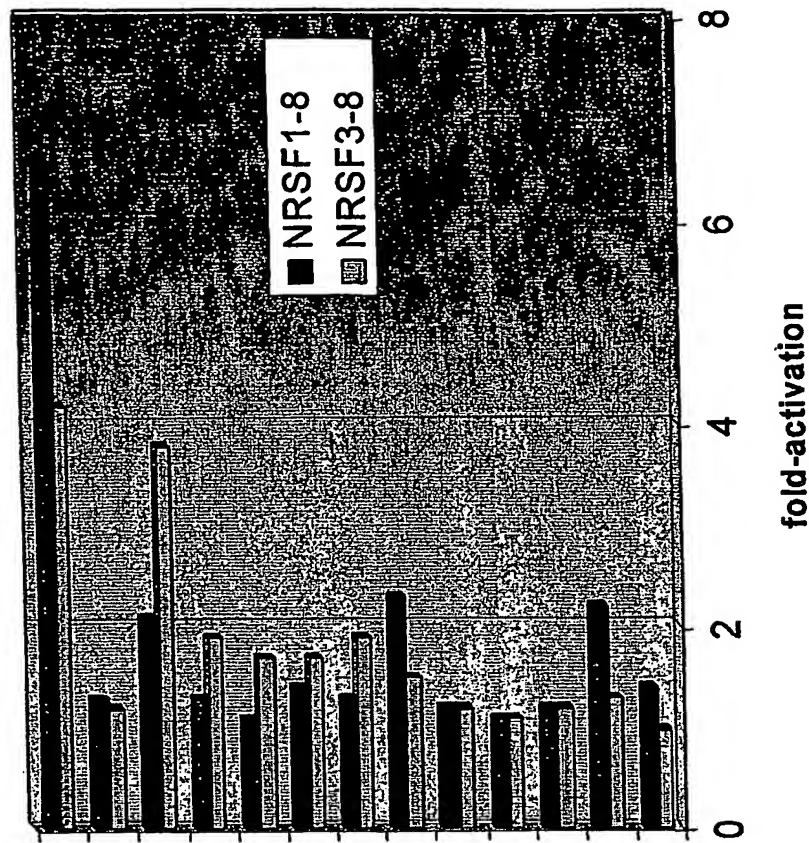


Figure 11

Figure13

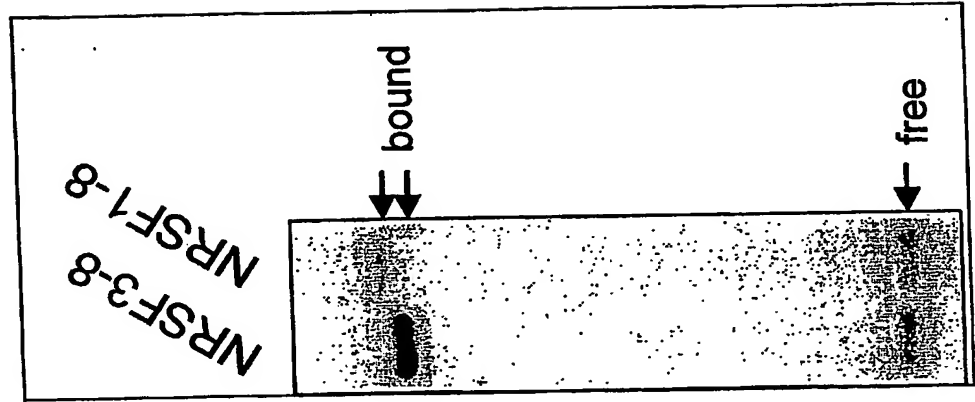
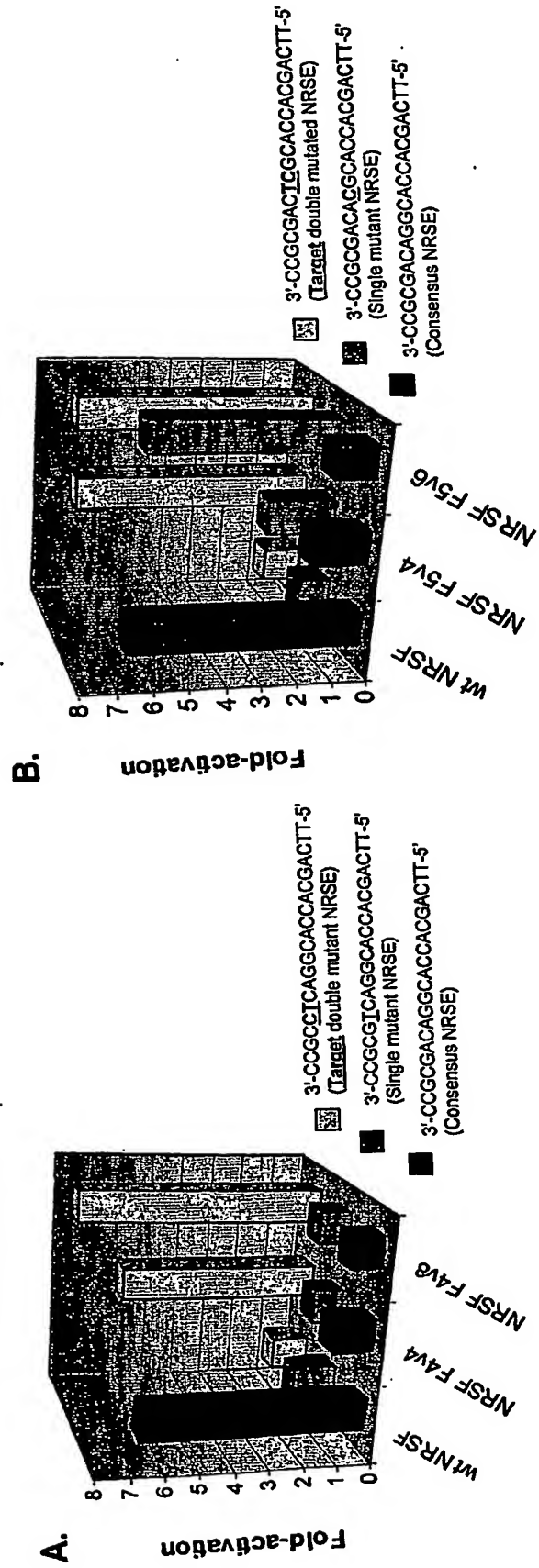


Figure 14

A. Finger 4 Selections		B. Finger 5 Selections	
3'-CCGCC <u>TC</u> AGGCACCACGACTT-5'		3'-CCGGC <u>ACT</u> CGCACCACGACTT-5'	
-1	1 2 3 5 6	-1	1 2 3 5 6
NRSF F4v1	H K T R M E	NRSF F5v1	T V G T R R
NRSF F4v2	H K T R M E	NRSF F5v2	T R G T K R
NRSF F4v3	H K T R M E	NRSF F5v3	T G S T R R
➡ NRSF F4v4	H R T R M E	➡ NRSF F5v4	T M S G R R
NRSF F4v5	H K T R K E	NRSF F5v5	T I S A R R
NRSF F4v6	H L T R K E	➡ NRSF F5v6	H M P T R R
NRSF F4v7	H K T R A E	NRSF F5v7	H R G T V R
➡ NRSF F4v8	H K T R D E	NRSF F5v8	R A P D K R

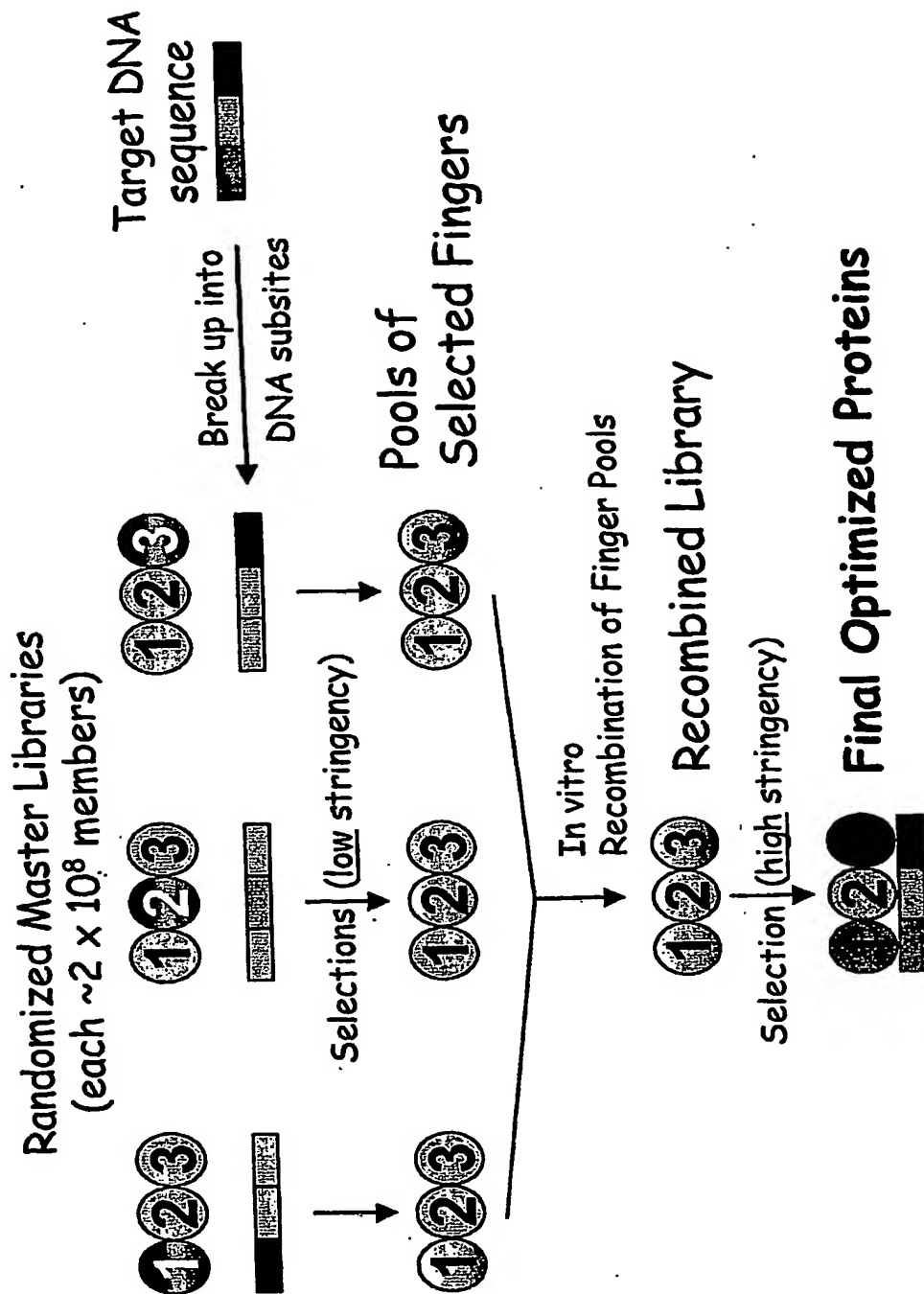
CCGGCCTCAGGCACCACGACTT-5'

Figure 15



bioRxiv preprint doi: <https://doi.org/10.1101/2017.04.11.128888>; this version posted April 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 16



115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

Figure 17

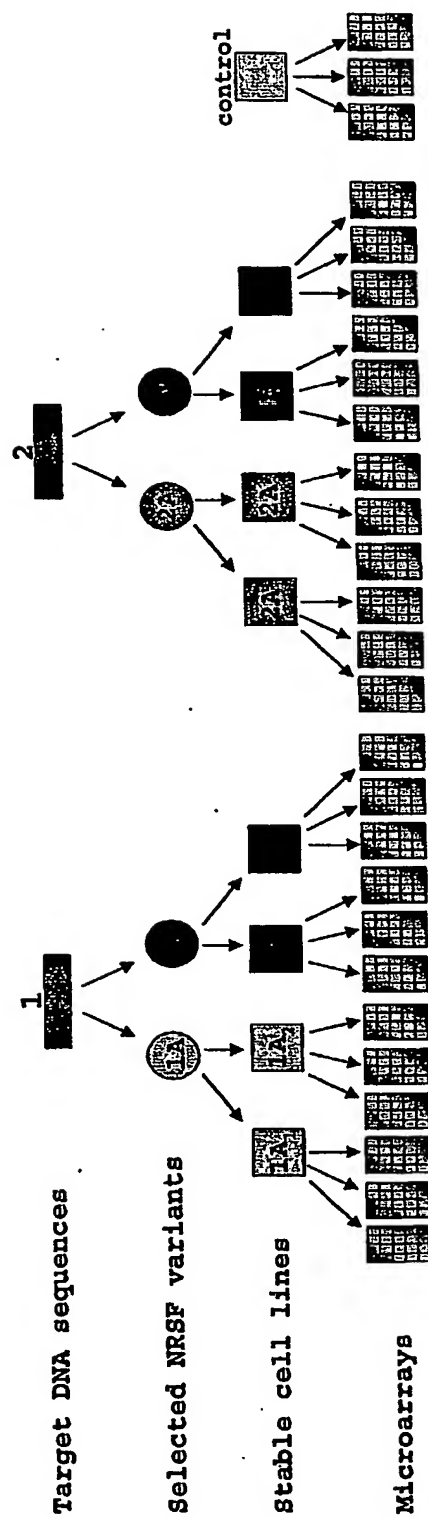


Figure 18

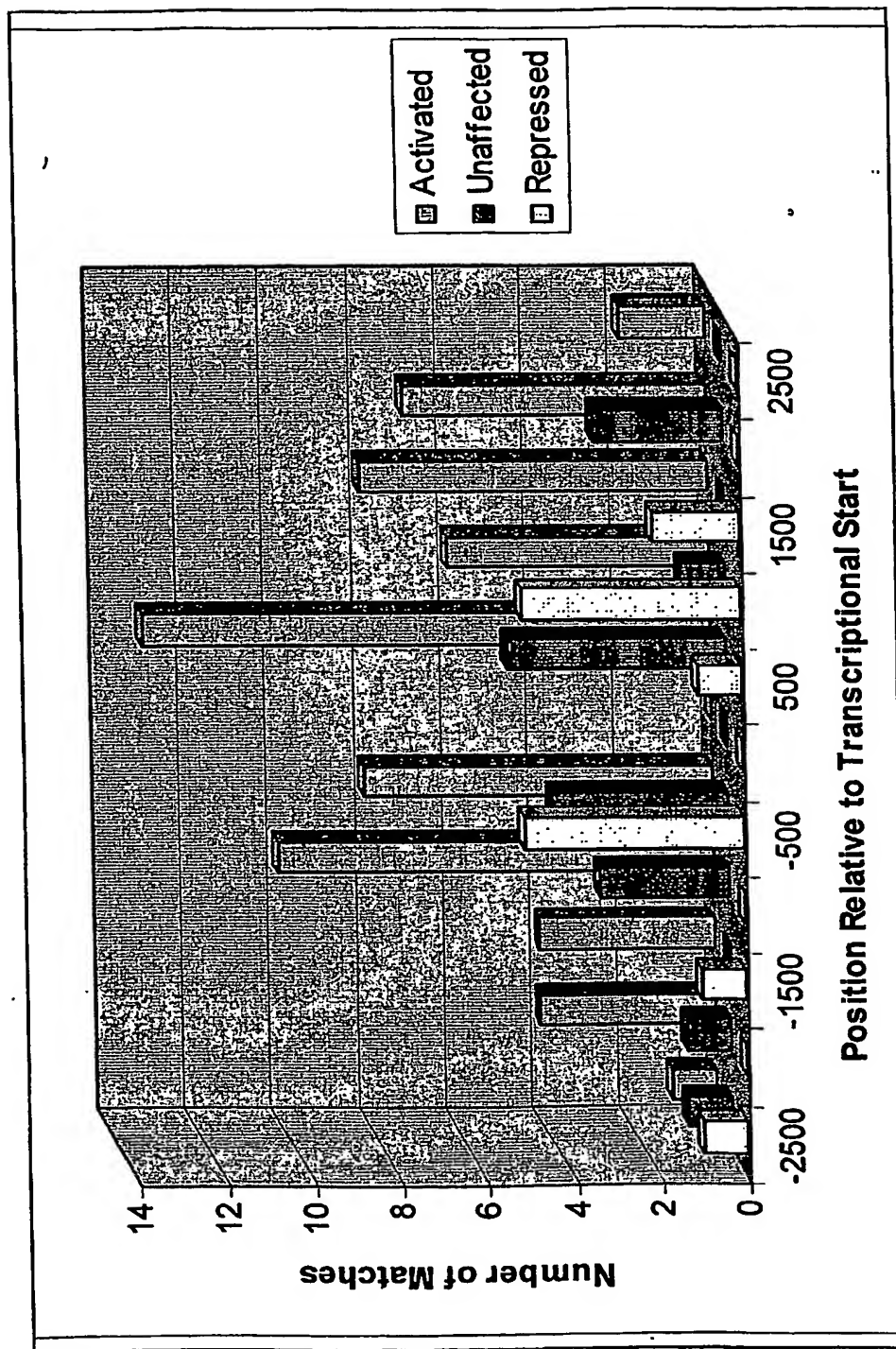


Figure 19

F4v1 (sequence identical to F4v2, F4v3)

MATQVMQSSGGGLFTSSGNIGMALPNDMYDLHDLSKAELAAPQLIMLANVALTGEVNGSCCDYLVGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLENNMELRSLSELSVVEPQPVFEASGAPDIYSSNKDLPPETPGAEDKGKSKTKPFR
CKPCQYEAEESEEQFVHHIRVHSAKFFVEESAQAKARESGSSTAEEGDFSKGPIRCDRCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCIICITYTTVSEYHWRKHLRNHFPRKVYTCGKCNVYFSDHKTRYMEHVRTHTGERPYKCELCPIYSSSQKTHLT
RHMRTSHGEKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPVCYAAASK
KCNLQYHFKSKHPTCPNKTMDVSKVKLKKTKKREADLPDNIITNEKTEIEQTKIKGDVAGKNEKSVKAEKRDVSKEKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGSNSEKFSKTKSKRKLKVDSHSLHGPVNDSEESTKTKKKVSKSKNNSQEVVK
GDSKVEENKQNTCMKKSTKKKTLKNKSSKKSSKPPQKEPVEKGSQAQMDPPQMGPAFTEAVQKGPVQVELPPMEHAQME
GAQIRPAPDEPVQMEVQEGPAQKELLPPVEPAQMVGAQIVLAHMEELPPPMETAQTEVAQMGPAFMEPAQMEVAQVESAP
MQVVQKEPVQMELSPPMEVQKEPVQIELSPPMVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVQKEPAQREPPPP
REPPLMHEPISKPPPLRKDKKEKSNMQSERARKEQVLIIEVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLNTGEGNKEAPLQKVGAEAEDELPGLAANINESTHISSSGQNLNTPGETLNGKHQTDIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPPSAVEEREAVSKTALASPPATMAANESQEIIDEDEGIHSHEGSDLSDNMSEGSDDSLGHGARVPVQ
ESSRKNAKEALAVKAAKGDVFCIFCDRSFRKGDYSKHLNRHLNVVYYLEEAAQGOE

Figure 20

于△节△

MATQVMGQSSGGGLFTSSGNI GMALPNDMYDLHDL SKAE LAAPQIIMLANVALTGEVNGSCCDYLVGEERQMAELMPVG
 DNNFSDSEEGEGLEESADIKGEPHGLENMEILRSLSVVEPQPVFEASGAPDIYSSNKDLPPETPGAEDKGKSSKTKKPF
 CKPCQYEAASEEQFVHHIRVHSAKFFVEESA EKQAKARESGSSTAEGDFSKGPIRCDRCGYNTNRYDHYTAHLKHHT
 AGDNERVYKCIICITYTTVSEYHWRKHLRNHFPRKVYTCGKCNFYFSDHRTRYMEHVRTHTGERPYKCELCIFYSSSQKTHLT
 RHMRTSHGEKPFKDCQCSYVASNOHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCVCDYAAASK
 KCNLQYHFKSKHPTCPNKTMDVSKVKLKKTKKREADLPDNI NEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKEKKPS
 NNVSIVQVTRTRKRSVTEVKEMDVHTGSNSEKFSKTKKSKRKLKLEVDSHSLHGPVNDSEESTKTKKKKVESKSKNNSQEVPK
 GDSKVEENKKQNTCMKSKTKKKTLLKNKSSKSSKPPQKEPVEKGSQAQMDPPQMGPA TEAVQKGPVQVELPPPMHAQME
 GAQIRPAPDEPVQMEVVQEGPAQKELLPPVEPAQMVGAQIVLAHME LPPMETAQTEVAQMGPA PMPAQMEVAQVESAP
 MQVVQKEPVQME LSPPMVEVVQKEPVQIELSPPMVEVVQKEPVQIELSPPIEVVQKEPVQME LSPPMGVVQKEPAQREPPPP
 REPLLHMEPI SKKPPLRKDKKEKSNMQSERARKEQVILEVGLVPVKDSWLLKESVSTEDLSPSPPLPKENLREEASGDQ
 KLLNTGEGNKEAPLQKVGAEEADESLPGLAANNINETHISSGQNLNTPGETLNGKHQTD SI VCEMKMDTDQNTRENLT
 GINSTVEEPVSPMLPPSAVEEREAVSKTALASPATMAANESQEIDEDEGIHSEGS DLSDNMSEGSDDSLHGARPVPQ
 ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLNVVYYLEAAQGOE

Figure 21

F4v5

MATQVMGQSSGGGLFTSSGNI GMALPNDWYDLHDL SKAELAAPQLIMIANVALTGEVNGSCCDYLVGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLENNMELRSLSELSVVEPQVFEASGAPDIYSSNKDLPPETPGAEDKGKSSKTKPFR
CKPCQYEAEESEEQFVHHIRVHSAKKFFVEESAQAKARESGSSTAEEGDFSKGPIRCDCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCICTYTTVSEYHWRKHLRNHFPRKVYTCGKCNYSFSDHKTRYKEHVRTHTGERPKCELCPYSSSQKTHLT
RHMRTSHGKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPVCDYAASK
KCNLQYHFKSKHPTCPNKTMDSVKVKKLTKKREADLPDNITNEKTEIEQTKIKGDVAGKNEKSVKAEKRDVSKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGSNSEKFSKTKSKRKLVDSSHSGPVNDEESSTKKKKVESKSNNSQEVPK
GDSKVEENKKQNTCMKKSTKKKTLKNKSSKSSKPPQKEPVEKGSQMDPPQMGPAPEAVQKGPVQVELPPMEHAQME
GAQIRPAPDEPVQMEVVQEGPAQKELLPPVEPAQMVGAIVLAHMELPPPMETAQTEVAQMGPAPEAVQMEVAQVESAP
MQVVQKEPVQMEVQMEVVQKEPVQIEVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
REPPLHMEPISKPPPLRKDKKESNMQSERARKEQVLIIEVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLLNTGEGNKEAPLQKVGAEEADESLPGLAANINESTHISSSGQNLNTPGETTLNGKHQTDTSIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQEI DEDEGIHSEGSDDLSDNMSEGSDDSLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLVNVYYLEEAAQGOE

Figure 22

F4v6

MATQVMGQSSGGGLFTSSGNIGMALPNDMYDLHDLASKAELAAPQLIMLANVALTGEVNGSCCDYLVGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLNEMELRSLSELSVVEPQPVFEASGAPDIYSSNKDLPPEPTGAEDKGKSSKTKPFR
CKPCQYEAEESEEQFVHHIRVHSAKFFVEESAQAKARESGSSTAEGDFSKGPIRCDCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCIICTYTTVSEYHWRKHLRNHFPRKVYTCGKCNYSFSDHLTRYKEHVTRHTGERPYKCELCPYSSSQKTHLT
RHMRTSHGKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCVCDYAASK
KCNLQYHFKSKHPTCPNKTMDSVKVCLKTKTKKREADLPDNIITNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGSNSEKFSKTKKSKRKLKLEVDSSHGHGPNDEESSTKKKKKVESKSKNNNSQEVPK
GDSKVEENKKQNTCMKKSTKKKTLKNKSSKSSKPPQKEPVEKGSQMDPPQMGPAFTEAVQKGPVQVELPPPMHAQME
GAQIRPAPDEPVQMEVQEGPAQKELLPPVEPAQMVGAQIVLAHMLPPPMETAQTEVAQMGPAPMEPAQMEVAQVESAP
MQVVQKEPVQMELSPPMEVVQKEPVQIELSPMEVVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVQKEPAQREPPPP
REPPLHMEPI SKKPPRLRKDKKESNMQSERARKEQVLEI EVGLVPVKDSWLLKESVSTEDLSPSPPLPKENLREEASGDQ
KLINTGEGNKEAPLQKVGAEEADESLPGLAANINESTHISSSGQNLNTPEGETLNGKHQTDIVCEMKMDTDQNTRENLT
GINSTVEEPPVSPMLPPSAVEEREAVSKTALASPATMAANESQEI DEDEGIHSHEGSDLSDNMSEGSDDSLGHGARPVPQ
ESSRKNVAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLVNVVYLEEAAQGE

00125291

Figure 23

F4v7

MATQVMQSSGGGLFTSSGNIGMALPNDMYDLHDL SKAELAAPQLIMLANVALTGEVNGSCCDYL VGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLENMELRSLELSVVEPQVFEASGAPDIYSSNKDLPDETPGAEDKGKSSKTKPFR
CKPCQYEAEESEEQFVHHIRVHSAKFFVEESAQAKARESGSSTAEGDFSKGPIRCDRCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCI ICTYTTVSEYHWRKHLRNHFPRKVYTCGKCNYSFSDHKT RYAEHVRTHTGERPYKCELCPYSSSQKTHLT
RHMRTSGEKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCVCDYAAASK
KCNLQYHFKSKHPTCPNKTMDVSKVKLKKTKKREADLPDNI TNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTTTRKSVTEVKEMDVHTGSNSEKFSKTKKSKRKL EVDSSHSLHGPVNDDEESSTK KKKKVESKKNNSQEVPK
GDSKVEENKQNTCMKKSTKKKTLKNKSSKKSSKPPQKEPVEKGS AQMDPPQMGPA TEAVQKGPVQVELPPMEHAQME
GAQIRPAPDEPVQMEVVEGPAQKELLPPVEPAQMVGAQIVLAHME LPPPMETAQTEVAQMGPAPMEPAQMEVAQVESAP
MQVVQKEPVQMELSPPMEVVQKEPVQIELSPMEVVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVVQKEPAQREPPPP
REPPLHMEPI SKKPPLRKDKKEKSNMQSERARKEQV LIEVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLNTGEGNKEAPLQKVGAEEADESLPGLAANINESTHIS SGGQNLNTPEGETLNGKHQ TDSIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQE IDEDEGIHSEGS DLSDNMSEGSDSLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLVNVVYYLEEAAQGGQE

Figure 24

F4v8

MATQVMQSSGGGLFTSSGNIGMALPNMDMYDLHDL SKAELAAPQLIMLANVALTGEVNGSCCDYL VGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEIPHGLENMELRSLLELSVVEPQPVFEASGAPDIYSSNKDLPDETPGAEDKGKSSKTKPFR
CKPCQYEAEESEEQFVHHIRVHSAKFFVEESA EKAKARESGSSTAEGDFSKGPIRCDRCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCI ICTYTTVSEYHWRKHLRNHFPRKVYTCGKCNYSFSDHKTRYDEHVRTHTGERPYKCELCYPYSSSQKTHLT
RHMRTHSGEKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTA DRSNFKKHVELHVNPRQFNCPVCDYAAASK
KCNLYHFKSKHPTCPNKTMDVSKVCLKTKTKKREADLPDNI TNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGNSNEKFSKTKKSKRKL EVDSHSLHGPVND EESSTK KKKVESKKNNSQEVPK
GDSKVEENKKQNTCMKKSTKKTLKNKSSKKSSKPPQKEPVEKGS AQMDPPQMGPAPTEAVQKGPVQVELPPPM EHAQME
GAQIRPAPDEPVQMEVVOEGPAQKELLPPVEPAQMVGAQIVLAHME LPPPMETAQTEVAQMGPAPEPAQMEVAQVESAP
MQVVQKEPVQMELSPPM EVVQKEPVQIELSPPM EVVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVVQKEPAQREPPPP
REPPLHMEPI SKKPPLRKDKKEKSNMQSERARKEQVLI EVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLLNTGEGNKEAPLQKVGAEEADESLPGLAANIN ESTHISSSGQNLNTPEGETLNGKHQTD SIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQEI DEDEGIHSEGS DLSDNMSEGSDDSGLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDY SKHLNRHLVNVVYILEEAAQQQE

00125291

Figure 25

F5V1

MATQVMGQSSGGGLFTSSGNI GMALPNDMYDLHLSKAELAAPQLIMLANVALTGEVNGSCCDYLVEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLENMELRSLSVVEPQPVFEASGAPDIYSSNKDLPPETPGAEDKGKSSKTKPFR
CKPCQYEAESSEEQFVHHIRVHSAKKFFVEESAQAKARESGSSTAEEGDFSKGPIRCDRCGYNTNRYDHYTAHLKHHTIR
AGDNERVYKCIICITYTTVSEYHWRKHLRNHFPRKVYTCGKCNFYSDRKNVYQHVTRHTGERPYKCELCPIYSSSTVGTLR
RHMRTSHGKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCDYAASK
KCNLQYHFKSKHPTCPNKTMDSVKVKKTKKREADLPDNI TNEKTEIEQTKIKGDVAGKKNESVKAERDVSKEKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGSNSEKFSKTKKSKRKL EVDSSHSLHGPVND EESSTK KKKKVESKKNNSQEVPK
GDSKVEENKKQNTCMKKSTKKKTLKNKSSKSSKPPQKEPVEKGSQAQMDPPQMGPPAPTEAVQKGPVQVELPPMEHAQME
GAQIRPAPDEPVQMEVVQEGPAQKELLPPVEPAQMVGAQIVLAHMELOPPMETAQTEVAQMGPPAPMEPAQMEVAQVESAP
MQVVQKEPVQMELSPPMEVVQKEPVQIELSPPEVVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVVQKEPAQREPPPP
REPPPLHMEPI SKKPPPLRKDKKEKSNMQSERARKEQVLEVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLINTGEGNKEAPLQKVGAEAEADESLPGLAANINETHISSSQNLNTPEGETLNGKHQTDIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQELDEDEGIHSEGS DLSDNMSEGSDDSGLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLVNVYYLEEAAQGGQ

Figure 26

F5v2

MATQVMGQSSGGGLFTSSGNI GMALPNDMYDLHDL SKAELAAPQLIMLANVALTGEVNGSCCDYL VGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLENMELRSL ELSVVEPQPVFEASGAPDIYSSNKDLP PETPGAEDKGKSSKTKPFR
CKPCQYEAESSEEQFVHHIRVHSAK KFFVEESA EKQAKARESGSSTAEEGDFSKGPI RCDRCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCI ICTYTTVSEYHWRKHLRNHFP RKVYTCGKCNFYFSDRKNNYVQHVRTHTGERP YKCELCPYSSSTRGTLK
RHMRTSGEKF FKCDCQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPVC DYAASK
KCNLQYHF KSKHPTCPNKTMDVSKV LKKT KKREADLPDNI TNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTTTRTRKSVTEVKEMDVHTGNSSEKFSKTKKSKRKL EVDSSH LGPVNDEESSTK KKKKVESKKNNSQEVPK
GDSKVEENKQNTCMKKSTK KKT LKNKSSKSKSPPOKEPVEKGSAQMDPPQMGPA PTEAVQKGPVQVELPPPM EHAQME
GAQIRPAPDEPVQMEV VQEGPAQKELLPPVEPAQMVGAQIVLAH MELPPP METAQTEVAQMGPA PMPAQMEVAQVESAP
MQVVKEPVQMELSPPMEV VQKEPVQIELSPMEV VQKEPVKIELSPPIEVVQKEPVQMELSPPMGV VQKEPAQREPPPP
REPPLHMEPI SKKPLR KDKKEKSNMQSERARKEQV LIEVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLINTGEGNKEAPLQKVGAEEADESLPGLAANIN ESTHISSSGQNLNTPEGETLNGKHQTD SIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQE IDEDEGIHSHEGSDLSDNMSEGSDDSGLHGARPVPQ
ESSRKNVAKALAVKAAKGFVCIFCDRSFRKGKDYSKHLNRHLNVVYYLEEAAQGGQ

Figure 27

F5V3

MATQVMGQSSGGGLFTSSGNIGMALPNDMYDLHDLSKAELAAPQLIMLANVALTGEVNGSCCDYLVGEERQMAELMPVG
DNNFSDSEEGELEESADIKGEPHGLENNMELRSLSVVEPQPVFEEASGAPDIYSSNKDLPPETPGAEDKGSSTKPPFR
CKPCQYEAEESEEFVHHIRVHSAKKFFVEESAEOAKARESGSSTAEEGDFSKGPIRCDRCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCIICTYTTVSEYHWRKHLRNHFPRKVYTCGKCNVFSDRKNVYVQHVRTHTGERPYKCELCPSYSSSTGSTLR
RHMRTSHGKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCVCDYAAASK
KCNLQYHFKSKHPTCPNKITMDVSKVKLKKTKKREADLPDNIITNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKEKKPS
NNVSVIQVTTTRTRKSVTEVKEMDVHTGSNSEKFSKTKSKRKLKVDSHSLHGPVNDSEESTKSKKVKESKKNNSQEVPK
GDSKVEENKQNTCMKKSTKKKTLKNKSSKSKPPQKEPVEKGSQAQMDPPQMGPAPEAVQKGPVQVELPPPMHAQME
GAQIRPAPDEPVQMEVQEGPAQKELLPPVEPAQMVGAQIVLAHMELEPPMETAQTEVAQMGPAPEVAQMEVAQVESAP
MQVVQKEPVQMELSPPMEVQKEPVQIELSPPMVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVVQKEPAQREPPPP
REPLHMEPI SKKPPLRKDKKEKSNMQSERARKEQVLEVGLVPVKDSWLLKESVSTEDLSPSPPLPKENLREEASGDQ
KLLNTGEGNKEAPLQKVGAEAEDESLPGLAANINESTHISSSGQNLNTPEGETLNGKHQTDIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPPSAVEEREAVSKTALASPPATMAANESQEI DEDEGIHSHGSDLSDNMSEGSDDSGLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGKDYSKHLNRHLVNVYYLEAAQGOE

Figure 28

F5v4

MATQVMQSSGGGLFTSSGNIGMALPNDMYDLHDLSKAELAAPQLIMLANVALTGEVNGSCCDYLVEERQMAELMPVG
DNNFSDSEEGEEESADIKGEPHGLENMELRSLSELSVVEPQPVFEASGAPDIYSSNKDLPPETPGAEDKGKSSKTKPFR
CKPCQYEAEESEEFVHHIRVHSAKFFVEESAQAKARESGSSTAEEGDFSKGPIRCDCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCIICITYTTVSEYHWRKHLRNHFPRKVYTCGKCNYSFSDRKNNYVQHVRTHTGERPYKCELCPYSSSTMSGLR
RHMRTSGEKPFFKCDQCSYVASNOHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCVCDYAASK
KCNLQYHFYSKHPTCPNKTMDSVKVKKTKTKKREADLPDNIITNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGSNSEKFSKTKKSKRKLKLEVDSHSLHGPVNDSESTKTKKKKVESKSKNNSQEVPK
GDSKVEENKKQNTCMKKSTKTKTLKNKSSKSSKPPQKEPVEKGSQAQMDPPQMGPPAPTEAVQKGPVQVELPPMEHAQME
GAQIRPAPDEPVQMEVVEGPAQKELLPPVEPAQMVGAQIVLAHMLPPPMETAQTEVAQMGPPAPMEPAQMEVAQVESAP
MQVVQKEPVQMELSPPMEVQKEPVQIELSPPMEEVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVVQKEPAQREPPPP
REPLHMEPI SKKPPPLRKDKKEKSNMQSERARKEQVLI EVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLINTGEGNKEAPLQKVGAEEADESLPGLAANINETHISSGQNLNTPEGETLNGKHQTDIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQEI DEDEGIHSEGSDDLSDNMSEGSDDSGLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLVNVYYLEEAAQGGQ

F5v5

Figure 30

F5v6

MATQVMQSSGGGLFTSSGNI GMALPNDMYDLHDLSKAELAAPQLIMLANVALTGEVNGSCCDYLVGEERQMAELMPVG
DNNFSDSEEGELEESADIKGEPHGLENMEIRSLSELVVEPQPVFEASGAPDIYSSNKDILPPETPGAEDKGKSSKTKPFR
CKPCQYEAESSEEQFVHHIRVHSAKFFVEESAQAKARESGSSTAEEGDFSKGPIRCDRCCYNTNRDHYHTAHLKHHIR
AGDNERVYKCIICITYTTVSEYHWRKHLRNHFPRKVTCTGKCNYSFSDRKNNYVQHVTRHTGERPYKCELCPIYSSSHMPTLR
RHMRTSGEKPFCDCQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCVCDYAAASK
KCNLQYHFKSKHPTCPNKTMDSVKVKKTKKREADLPDNIITNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGSNSEKFSKTKKSKRKLVDSDSHSLHGPVNDDEESSTKSKKVESKSKNNSQEVPK
GDSKVEENKQNTCMKKSTKKKTLKNKSSKKSSKPPQKEPVEKGSQAQMDPPQMGPA TEAVQKGPVQVELPPPEHAQME
GAQIRPAPDEPVQMEVVQEGPAQKELLPPVEPAQMVGAQIVLAHME LPPPMETAQTEVAQMGPAPEVAQVESAP
MQVVQKEPVQMELSPPMEEVVQKEPVQIELSPMEVVQKEPVKIELSPPIEVVQKEPVQMELSPPMGVVQKEPAQREPPPP
REPLHMEPI SKKPPLRKDKKEKSNMQSERARKEQV LIEVGLVPVKDSWLLKESVSTEDLSPSPPLPKENLREEASGDQ
KLLNTGEGNKEAPLQKVGAEEADESLPGLAANINESTHISSSGQNLNTPGETLNGKHQTD SIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQEI DEDEGIHSEGSDDLSDNMSEGSDDSGLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLNVNYYLEEAQQQE

Figure 31

F5v7

MATQVMQSSGGGLFTSSGNIGMALPNDMYDLHDL SKAELAAPQLIMLANVALTGEVNGSCCDYLVGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLENMELRSLELSVPEPQPVFEASGAPDIYSSNKDLPPETPGAEDKGKSKTKPFR
CKPCQYEAASEEQFVHHIRVHSAKKFFVEESAQAKARESGSSTAEEGDFSGPIRCDRCGYNTNRYDHYTAHLKHHTR
AGDNERVYKCIICCTYTTVSEYHWRKHLRNHFPRKVTGKCNYPFSDRKNNYVQHVRTHTGERPYKCELCPYSSSHRGTLV
RHMRTSGEKPFKCDQCSYVASNQHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPVCDYAASK
KCNLQYHFKSKHPTCPNKTMDVSKVLKKTKKREADLPDNI TNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTTTRTRKSVTEVKEMDVHTGSNSEKFSKTKKSKRKLKVDSHSLHGPVNDDEESTKKKKVSKSKNNSSQEVPK
GDSKVEENKKQNTCMKKS TKKTLKNKSSKSSKPPQKEPVEKSSAQMDPPQMGPA PTEAVQKGPVQVELPPPMHAQME
GAQIRPAPDEPVQMEVVQEGPAQKELLPPVEPAQMVGAQIVLAHME LPPPMETAQTEVAQMGPA PMPAQMEVAQVESAP
MQVVQKEPVQMELSPPMEVVQKEPVQIELSPPMVEVVQKEPVQIELSPPIEVVQKEPVQMELSPPMGVVQKEPAPPPPP
REPLHMEPI SKKPPLRKDKKEKSNMQSERARKEQVLIEVGLVPVKDSWLLKESVSTEDLSPPSPPLPKENLREEASGDQ
KLINTGEGNKEAPLQKVGAEEADESLPGLAANINESTHISSSGQNLNTPGETLNGKHQTD SIVCEMKMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQEI DEDEGIHSHEGSDLSDNMSEGSDDSGLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCIFCDRSFRKGDYSKHLNRHLVNVYYLEEAAQGE

Figure 32

F5v8

MATQVMGQSSGGGLFTSSGNI GMALPNDMYDLHDL SKAELAAPQLIMLANVALTGEVNGSCCDYLVGEERQMAELMPVG
DNNFSDSEEGEGLEESADIKGEPHGLENMELRSLLELSVVEPQPVFEASGAPDIYSSNKDLPPETPGAEDKGKSSKTKPFR
CKPCQYEAESSEEQFVHHIRVHSAKFFVEESA EKQAKARESGSSTAEEGDFSKGPIRCDRCCYNTNRYDHYTAHLKHHTR
AGDNERVYKCI ICTYTTVSEYHWRKHLRNHFPRKVYTCGKCNYSFDRKNNYVQHVRTHTGERPYKCELCPIYSSSRAPDLK
RHMRTSHGEKPFKCDQCSYVASNOHEVTRHARQVHNGPKPLNCPHCDYKTADRSNFKKHVELHVNPRQFNCPCVDYAAASK
KCNLQYHFKSKHPTCPNKTMDVSKVKLKKTKKREADLPDNI TNEKTEIEQTKIKGDVAGKKNEKSVKAEKRDVSKKKPS
NNVSVIQVTRTRKSVTEVKEMDVHTGSNSEKFSKTKKSKRKLEVDSSHSLHGPVNDSEESTKKKKVESKSKNNSQEVPK
GDSKVEENKQNTCMKKSTKKKTLKNKSSKSSKPPQKEPVEKGS AQMDPPQMGPA TEAVQKGPVQVELPPFMEHAQME
GAQIRPAPDEPVQMEVVQEGPAQKELLPPVEPAQMVGAQIVLAHME LPPPMETAQTEVAQMGPAPMEPAQMEVAQVESAP
MQVVQKEPVQME LSPPMEEVVQKEPVQIELSPPMEEVVQKEPVQIELSPPIEVVQKEPVQME LSPPMGVVQKEPAQREPPPP
REPPLHMEPI SKKPPLRKDKKEKSNMQSERARKEQVLIEVGLVPVKDSWLLKESVSTEDLSPSPPLPKENLREEASGDQ
KLINTCEGNKEAPLQKVGAEEADESLPGLAANINESTHISSSGQNLTPEGETLNGKHQTD SIVCEMKVMDTDQNTRENLT
GINSTVEEPVSPMLPPSAVEEREAVSKTALASPPATMAANESQEIDEDEGIHSGSDLSDNMSEGSDDSGLHGARPVPQ
ESSRKNAKEALAVKAAKGDFVCI FCDRSFRKGDYSKHLNRHLVNVVYYLEEAAQCQE

Figure 33

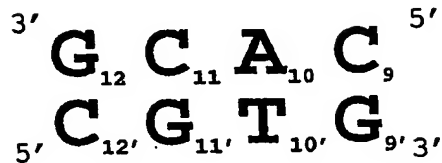
A.

wild-type F6

-112356

NQHETR

relevant portion
of NRSE



B.

F6 variants/
base 9

-112356

DRGNRR

DRGNNR

DKANAR

DLSNRR

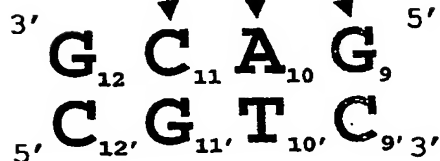
DSTNKR

ERGNQR

ERYAVR

EKYKVI

D+ N R



F6 variants/
base 11

-112356

RREREL

RREKVM

RRERYI

RRDNET

RRDGAN

RKDLAL

RADIRL

RLELKV

RRD

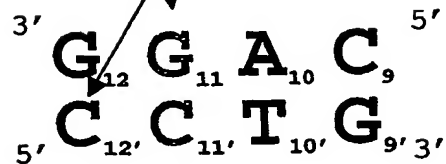
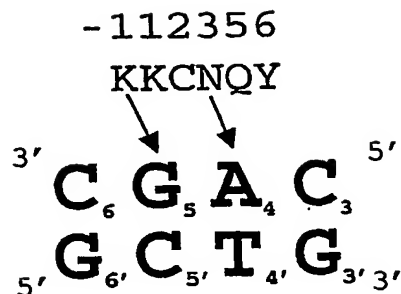


Figure 34

A.

wild-type F8

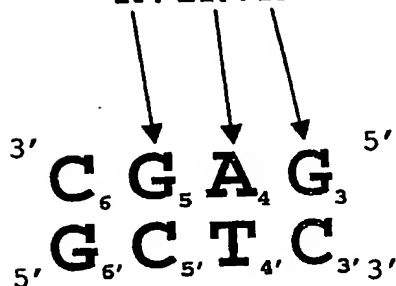
relevant portion
of NRSE



B.

F8 variants/
base 3

-112356
RLFNRR
KKYNRR
RQYNQR
TKFNHR
RKYNRR
RKYNRR
RRANVR
RFYNRR
R+YN+R



F8 variants/
base 4

-112356
RRSTRY
RRSTRY*
RRSTRY*
RKATDY
RRTTLY
RKATMY
RRSTQY
RRSTVY
R+sT_Y

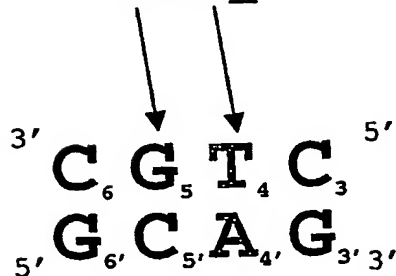
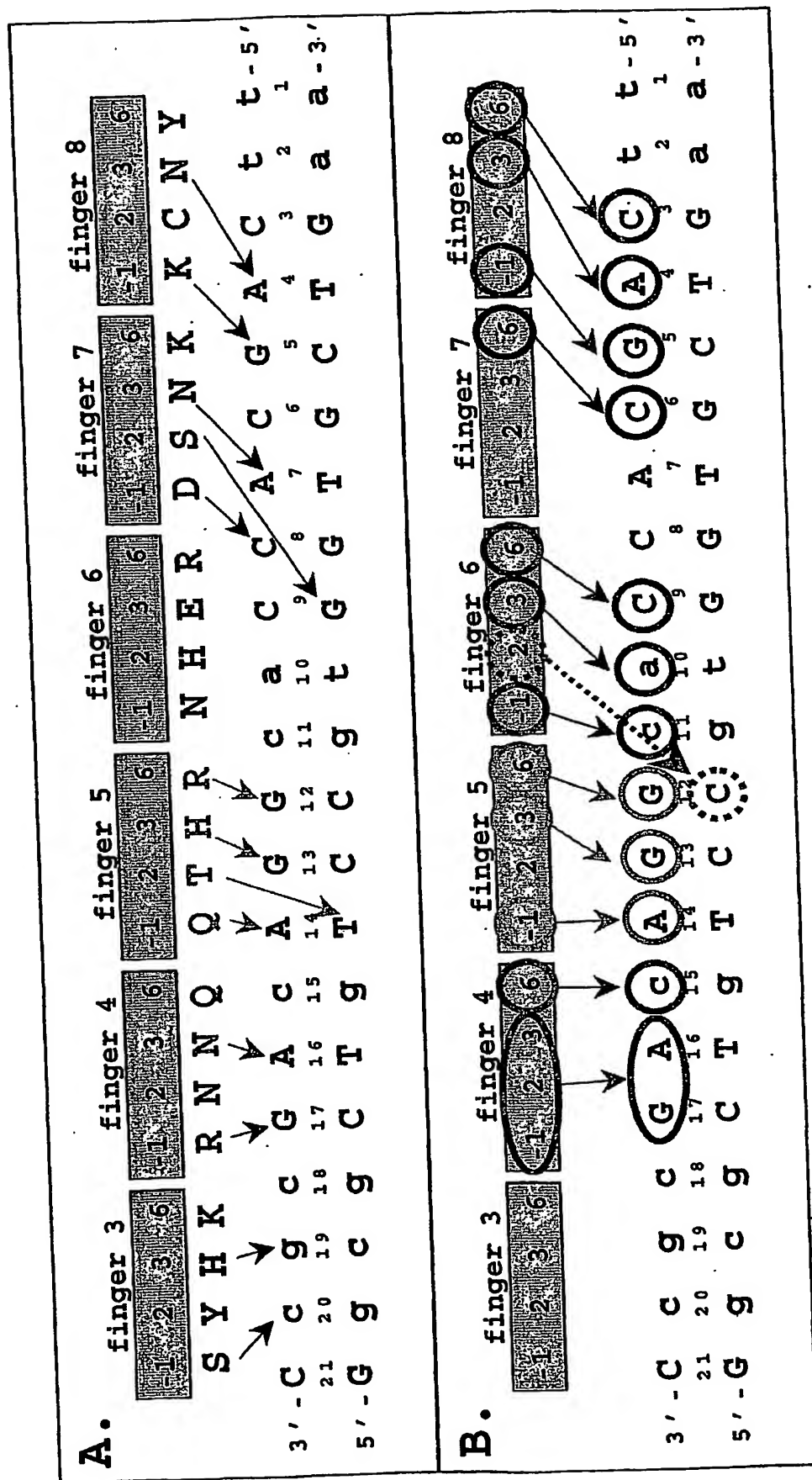


Figure 35



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.